第 39 卷　第 7 期　2015 年 4 月 10 日
DOI：10.7500/AEPS20140111003
电力系统自动化
Automation of Electric Power Systems
Vol.39 No.7 Apr.10,2015

基于时间序列分析的输变电设备状态大数据清洗方法

严英杰[1]，盛戈皞[1]，陈玉峰[2]，江秀臣[1]，郭志红[2]，秦少鹏[3]
（1. 上海交通大学电子信息与电气工程学院，上海市 200240；2. 国网山东省电力公司电力科学研究院，山东省济南市 250002；3. 国网四川省电力公司广安供电公司，四川省广安市 638500）

摘要：数据清洗是输变电设备状态评估数据预处理的一个关键步骤，有助于提高数据质量和数据利用率。文中将设备状态信息等效成各状态量的时间序列，提出了一种基于时间序列分析的双循环迭代检验法。首先，将时间序列中的异常数据进行了分类，并将缺失值归纳为其中一类异常值。然后，分析了不同类别异常值对时间序列模型的影响，并阐述了迭代检验法的实现步骤。最后，利用所述方法对南网某变压器和线路的监测数据进行了数据清洗，结果表明该方法能识别并修正数据中的噪声点，填补缺失值，满足数据清洗要求。

关键词：大数据；数据清洗；时间序列；电力设备状态数据

0　引言

传统的设备评估诊断大多基于单一部件、单一参量的阈值判断，由于设备测试手段的局限性、故障机理的复杂性、运行环境的多样性、知识的不精确性导致诊断评价结果片面、缺少故障发展全面分析和预测的手段等问题[1-2]。对设备在线监测、带电检测、离线试验等设备全景状态信息进行状态检测，提升输变电设备评价与异常诊断的准确性是设备状态评估诊断技术的发展趋势[3-6]。

大数据是目前学术界和产业界共同关注的研究主题，具有广阔的应用前景。随着电力系统的发展，电力设备在线监测数据及生产管理、运行调度等数据逐步在统一的信息平台上完成集成共享，为大数据技术融合输变电设备状态数据的分析处理创造了条件。目前大数据技术在电力行业中的应用主要集中在电网大数据的传输和存储[7-8]及电力负荷数据的分析处理上[9-11]。

输变电设备全景状态信息呈现来源多、信息异构、数量庞大、属性繁多等特点，其数据往往是不完整的、有噪声的和不一致的。状态量原始的数据质量往往不能满足后续状态评价模型的要求，因此，在状态评估或诊断分析之前进行数据清洗是必不可少的。数据清洗通过填充缺失值、平滑噪声数据和识别离群点来提高数据质量，有助于提高数据挖掘过程的准确率和效率[12-13]。

在输变电设备数据清洗方面，国内外的研究如文献[14-16]所示。文献[14]在建立故障与信息的映射关系时将海量数据通过粗糙集信息熵的方法进行了约简，从而解决了数据缺失的问题，但是破坏了数据自身信息的完整性。文献[15-16]在处理支持向量机训练集的噪声和异常数据时使用了模糊 C 均值聚类方法，通过计算数据到聚类中心的距离来分离出噪声数据。但是这种聚类方法将分离出的噪声数据直接剔除，破坏了状态量数据链的连续性。以上研究在数据清洗过程中造成了数据的丢失，不利于在后续状态评估中对数据本身信息的挖掘。

本文提出了一种基于时间序列分析的数据清洗方法，其原理是利用时间序列模型识别各状态量的时间序列，检测出数据的异常模式，判断异常数据是能提取设备故障信息的“有用数据”还是可被清洗的“无用数据”。当异常数据是由设备异常状态产生时，用时间序列干预模型进行拟合以提取有效故障信息。在数据清洗时，根据序列中异常值的种类选择不同的修正公式，从而达到修正噪声点数据和填补缺失值的目的。相比于传统的删除噪声点，本文方法清洗出的数据是不带有噪声点和缺失值的数据，从而避免了时间序列中有用信息的丢失，更能有效地反映原始时间序列的动态变化，适应输变电设备状态数据的特点。

收稿日期：2014-01-11；修回日期：2014-09-02。
国家自然科学基金资助项目(51477100)；国家高技术研究发展计划(863 计划)资助项目(SS2012AA050803)；国家电网公司科技项目。

1　基于时间序列的输变电设备状态数据清洗方法原理

1.1　状态数据的特点及时间序列方法适用性

输变电设备状态量的检测是由各个传感器来完成的，但是经过底层的预处理而上传到数据库进行状态评估的原始数据可以认为是按时间序列排列的特征量数据。这些数据的统一格式为“时间.特征量＝数值”，因此，可认为采集的所有状态量形成了一个单元或多元的连续而完整的时间序列[17]，如矩阵 $\boldsymbol{X}$ 所示：

$$\boldsymbol{X}=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1h} \\ X_{21} & X_{22} & \cdots & X_{2h} \\ \vdots & \vdots & & \vdots \\ X_{l1} & X_{l2} & \cdots & X_{lh} \end{bmatrix} \tag{1}$$

式中：X_{lh} 为在 h 时刻状态量 l 的数值。

输变电设备正常运行状态下的状态数据一般呈现如下 3 种规律，并都可适用于时间序列方法：①状态量幅值变化较小，如导线拉力、接地电流、油中气体 C_2H_2 等，这些状态量数据都属于平稳序列，可直接用自回归移动平均函数 ARMA(p,q)拟合；②状态量呈缓慢上升趋势，如油中气体 CO 和 CO_2，可以通过差分方法转化为平稳序列，并用自回归求和移动平均函数 ARIMA(p,d,q)拟合；③状态量呈周期性变化，在时间序列上表现为 s 个时间间隔后的观测点呈现相似性，如油温、导线温度等，可通过 ARIMA(p,d^s,q)拟合。

根据输变电设备的运行特点，状态数据中的异常通常表现为两种形式：①可用于数据清洗的异常，即噪声点和缺失值；②设备运行状态受到干扰而导致的数据异常。噪声点是指由于仪器异常或设备系统的扰动引起的严重偏离期望值的数据，这些数据不仅会影响模型拟合的精度，而且会导致后续状态评估出现偏差，引起误诊。缺失值是指由于传感器的短时失效、通信端口异常、记录失误等因素引起的数据中断，状态数据中存在的缺失值破坏了系统运行的连续性，不利于后续的状态评估和趋势检验。设备在运行过程中会产生突发性故障、绝缘劣化等，这些常常会引起数据的水平迁移异常和趋势改变性异常，此类异常数据反映了设备运行工况的异常，不属于清洗范畴。设备状态数据的时间序列中往往含有多个异常数据，修复所有的噪声点和缺失值是设备状态数据清洗的目标，同时也要实现突发性故障信息的有效获取，而不是作为异常数据剔除。

1.2　可用于清洗的异常数据

1.2.1　噪声点和缺失值的模型分类

时间序列中的噪声点可以分为新息异常值(IO)、附加异常值(AO)和两种类型异常值的组合[18]。设 X_t 是无异常值的时间序列，X_t 服从 ARIMA(p,d,q)，可表示为：

$$X_t=\frac{\theta(B)}{\varphi(B)\nabla^d}a_t \tag{2}$$

$$\theta(B)=1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q \tag{3}$$

$$\varphi(B)=1-\varphi_1B-\varphi_2B^2-\cdots-\varphi_qB^q \tag{4}$$

式中：B 为延迟算子；$\theta(B)$和 $\varphi(B)$分别为没有公共因子的平稳和可逆算子；$\theta_1,\theta_2,\cdots,\theta_q$ 为 $\theta(B)$的相应参数；$\varphi_1,\varphi_2,\cdots,\varphi_q$ 为 $\varphi(B)$的相应参数；a_t 为相互独立，具有相同分布 $N(0,\sigma_a^2)$的白噪声序列，其中 σ_a 为含异常值的残差的标准差；$\nabla=1-B$，适用于 1.1 节中符合第 2 和第 3 种规律的状态数据(即具有趋势性、周期性的时间序列)。

用 Z_t 表示观测到的时间序列，那么 T 时刻(脉冲发生时刻)包含噪声点的 ARIMA(p,d,q)可表示为以下 3 种噪声点模型。

1)IO 模型

$$Z_t=X_t+\omega\frac{\theta(B)}{\varphi(B)\nabla^d}I_t^{(T)}=\frac{\theta(B)}{\varphi(B)\nabla^d}(a_t+\omega I_t^{(T)}) \tag{5}$$

$$I_t^{(T)}=\begin{cases}1 & t=T\\ 0 & t\neq T\end{cases} \tag{6}$$

式中：ω 为异常值影响因子；$I_t^{(T)}$ 为脉冲函数。

IO 影响了 T 时刻之后的所有观测值 Z_T，Z_{T+1},…。其影响效应与 Z_t 的模型形式有关，通过 $\theta(B)/\varphi(B)$所描述的系统动态特性而影响后面的所有观测序列。

2)AO 模型

$$Z_t=X_t+\omega I_t^{(T)}=\frac{\theta(B)}{\varphi(B)\nabla^d}a_t+\omega I_t^{(T)} \tag{7}$$

AO 只影响该干扰发生的那一时刻 T 上的序列值，而不影响该时刻以后的序列值。AO 通过未知的 ω 而起作用。时间序列中的缺失值可以认为是一种 AO。

3)多个异常值的模型

在通常情况下，一个被观测的时间序列可以在不同的时间点上受不同类型的异常值的影响，因此，得到下面两种异常值组合的模型：

$$Z_t=X_t+\sum_{j=1}^{k}\omega_jv_j(B)I_t^{(T)} \tag{8}$$

$$v_j(B)=\begin{cases}\dfrac{\theta(B)}{\varphi(B)\nabla^d} & \text{IO}\\ 1 & \text{AO}\end{cases} \tag{9}$$

式中：k 为异常值个数；ω_j 和 v_j 分别为对应于不同异常值的影响因子和算子。

1.2.2　异常数据对时间序列拟合的影响

异常数据会影响时间序列拟合的精度，通过对拟合残差的分析可以将两类异常数据的影响量化。设时间序列拟合的残差为 e_t，则

$$e_t=\pi(B)Z_t \tag{10}$$

$$\pi(B)=\frac{\theta(B)}{\varphi(B)\nabla^d}=1-\pi_1B-\pi_2B^2-\cdots \tag{11}$$

式中：$\pi(B)$为表征残差影响的算子；$\pi_1,\pi_2,\cdots$为$\pi(B)$的相应参数。

在观测到的时间序列 Z_t 中存在异常数据时，拟合残差序列 e_t 可以表示为：

$$e_{t,\mathrm{AO}}=\omega\pi(B)I_t^{(T)}+a_t \tag{12}$$

$$e_{t,\mathrm{IO}}=\omega I_t^{(T)}+a_t \tag{13}$$

式(12)和式(13)分别表示了异常数据为 AO 和 IO 时，拟合残差序列与白噪声序列的关系。将式(12)用矩阵的方式扩展开来，对长度为 n 的时间序列，式(12)可写为：

$$\begin{bmatrix}e_{1,\mathrm{AO}}\\ \vdots\\ e_{T-1,\mathrm{AO}}\\ e_{T,\mathrm{AO}}\\ e_{T+1,\mathrm{AO}}\\ \vdots\\ e_{n,\mathrm{AO}}\end{bmatrix}=\omega\begin{bmatrix}0\\ \vdots\\ 0\\ 1\\ -\pi_1\\ \vdots\\ -\pi_{n-T}\end{bmatrix}+\begin{bmatrix}a_1\\ \vdots\\ a_T\\ a_{T+1}\\ a_{T+2}\\ \vdots\\ a_n\end{bmatrix} \tag{14}$$

由于 a_t 是白噪声序列，根据式(14)由最小二乘理论算得噪声点 AO 对时间序列拟合的影响 $\hat{\omega}_{\mathrm{AO}}$ 为：

$$\hat{\omega}_{\mathrm{AO}}=\frac{e_{T,\mathrm{AO}}-\sum_{i=1}^{n-T}\pi_i e_{T+i,\mathrm{AO}}}{\sum_{i=0}^{n-T}\pi_i^2} \tag{15}$$

同理，噪声点 IO 对时间序列拟合的影响 $\hat{\omega}_{\mathrm{IO}}$ 为：

$$\hat{\omega}_{\mathrm{IO}}=e_{T,\mathrm{IO}} \tag{16}$$

因此，在时刻 T，IO 对时间序列拟合影响的最好量化估计是残差 $e_{T,\mathrm{IO}}$，而 AO 影响的最好量化估计是残差 $e_{T,\mathrm{AO}},e_{T+1,\mathrm{AO}},\cdots,e_{n,\mathrm{AO}}$的线性组合，其权数依赖于时间序列的结构。

1.2.3　异常数据的检验统计量

时间序列中异常值的存在将使得参数估计产生严重的偏差，这些偏差根据 1.2.2 节中噪声点 AO 和 IO 对时间序列拟合的影响 $\hat{\omega}_{\mathrm{AO}}$ 和 $\hat{\omega}_{\mathrm{IO}}$，可以综合成噪声点的检验统计量，当检验统计量超过一定的限值时，可以判断其对应的时刻 T 存在噪声点。每个观测点的 AO 和 IO 的检验统计量如下：

$$T_{\mathrm{IO}}^{t'}=\frac{\hat{\omega}_{\mathrm{IO}}}{\sigma_a} \tag{17}$$

$$T_{\mathrm{AO}}^{t'}=\frac{\hat{\omega}_{\mathrm{AO}}}{\sigma_a}\sqrt{\sum_{i=T}^{n}\pi_i^2} \tag{18}$$

式中：t'为异常数据产生的时刻；$T_{\mathrm{AO}}^{t'}$为 AO 的检验统计量，$T_{\mathrm{IO}}^{t'}$为 IO 的检验统计量，两者的极限分布均为标准正态分布。

1.3　反映设备状态的异常数据

通过对输变电设备突发性故障的统计分析[19-20]可知，故障时其状态数据往往会产生水平迁移和快速变化的趋势，这种情况下状态数据用式(1)拟合时在某一时间点后的残差序列均远大于之前的值，因此，可直接判断数据不可做清洗，只能通过时间序列干预模型拟合。状态数据的两种干预响应结构如下。

1)反映水平迁移的干预响应结构为：

$$\omega B^bS_t^{(T)} \tag{19}$$

$$S_t^{(T)}=\begin{cases}0 & t<T\\ 1 & t\geqslant T\end{cases} \tag{20}$$

式中：$S_t^{(T)}$ 为阶跃函数；b 为延迟时间。

该结构说明输入的干预变量是 $S_t^{(T)}$，输出的状态量延迟 b 后做出反应且强度为 ω，以后再不回到以前的状况。这类干预影响反映出了状态量的水平迁移，如变压器对地绝缘故障时铁芯接地电流迅速变大而超过限值(100 mA)等。

2)反映趋势改变的干预响应结构为：

$$\frac{\omega B^b}{1-\delta B}S_t^{(T)} \tag{21}$$

式中：δ 为延迟算子的相应参数。

此类干预影响常常用来表示趋势性状态量趋势的变化。如反映变压器固体绝缘的 CO/CO_2，在正常情况下其数值是缓慢上升的，当变压器固体绝缘受到破坏而导致劣化加速时，CO 数值会呈快速上升趋势，时间序列的斜率比正常情况下大很多。在对 CO 的时间序列做一阶差分后符合该类干预影响结构。

2　输变电设备状态信息数据清洗步骤

设备状态信息获取方式的多样性及采集间隔的不确定性使得各状态量时间序列的参数是未知的、异常数据产生的时刻 T 是不确定的，因此，时间序列模型的搭建、模型参数估计、异常数据类型识别是必不可少的数据清洗步骤。由于异常数据的存在将使时间序列参数的估计产生偏差，因此，针对噪声点出现时刻与个数未知、预先没有模型参数的情况，使

用迭代检验的方法对观测的时间序列进行数据清洗,共分为7个步骤(流程图见附录A图A1)。

步骤1:假定不存在异常值,对观测序列 Z_t 建立时间序列模型,并由所估计的模型计算初始残差,即

$$\hat{e}_t=\hat{\pi}(B)Z_t=\frac{\hat{\varphi}(B)\nabla^d}{\hat{\theta}(B)}Z_t \tag{22}$$

式中:$\hat{e}_t$ 为初始拟合的残差序列;$\hat{\pi}(B)$ 为 $\pi(B)$ 的初始值;$\hat{\theta}(B)$ 和 $\hat{\varphi}(B)$ 分别为初始拟合的平稳和可逆算子。

残差方差的初始估计 $\hat{\sigma}_a^2$ 为:

$$\hat{\sigma}_a^2=\frac{1}{n}\sum_{t=1}^{n}\hat{e}_t^2 \tag{23}$$

步骤2:观测拟合残差序列。若从某时间点开始残差序列呈现水平迁移,并远大于之前的残差值,则原始时间序列需用干预模型拟合,跳至步骤7;否则跳至外循环。

步骤3:在外循环中,利用已估计的模型,对 $t=1,2,\cdots,n$,计算每个观测点的检验统计量 T_{AO}^t 和 T_{IO}^t。

定义 $\lambda_{T_{\max}}=\max\{|T_{\mathrm{AO}}^t|,|T_{\mathrm{IO}}^t|\}$,这里 $T_{\max}$ 为最大值发生的时刻。当 $\lambda_{T_{\max}}>C$ 时,其中 C 是预先确定的常数,通常取3和4之间的值,则说明存在异常数据,进入内循环修正数据。

步骤4:在内循环中修正数据。

当 $\lambda_{T_{\max}}=|T_{\mathrm{AO}}^{T_{\max}}|>C$ 时,可以确定在时刻 $T_{\max}$ 存在异常数据AO,其对模型拟合的影响 $\hat{\omega}_{\mathrm{AO}}$ 通过式(15)可以求得。通过式(7)修正原始时间序列数据,得到新的时间序列 $\widetilde{Z}_t$ 为:

$$\widetilde{Z}_t=Z_t-\hat{\omega}_{\mathrm{AO}}I_t^{(T)} \tag{24}$$

并由式(12)修正得到新的残差 $\widetilde{e}_{t,\mathrm{AO}}$ 为:

$$\widetilde{e}_{t,\mathrm{AO}}=\hat{e}_t-\hat{\omega}_{\mathrm{AO}}\hat{\pi}(B)I_t^{(T)} \tag{25}$$

当 $\lambda_{T_{\max}}=|T_{\mathrm{IO}}^{T_{\max}}|>C$ 时,确定在时刻 $T_{\max}$ 存在异常数据IO,其对模型拟合的影响 $\hat{\omega}_{\mathrm{IO}}$ 可通过式(16)求得,利用式(5)修正数据,则IO的影响可以消除,即

$$\widetilde{Z}_t=Z_t-\frac{\hat{\theta}(B)}{\hat{\varphi}(B)\nabla^d}\hat{\omega}_{\mathrm{IO}}I_t^{(T)} \tag{26}$$

并由式(13)修正得到新的残差 $\widetilde{e}_{t,\mathrm{IO}}$ 为:

$$\widetilde{e}_{t,\mathrm{IO}}=\widetilde{e}_t-\hat{\omega}_{\mathrm{IO}}I_t^{(T)} \tag{27}$$

使用迭代的方法识别并修正时间序列所有的噪声点。在修正后的残差 $\widetilde{e}_{t,\mathrm{AO}}$,$\widetilde{e}_{t,\mathrm{IO}}$ 和残差标准差 $\widetilde{\sigma}_a^2$ 的基础上再次计算每个观测点的检验统计量 $T_{\mathrm{AO}}^{t'}$ 和 $T_{\mathrm{IO}}^{t'}$,并重复步骤4,直到所有的异常数据都被识别出来。当 $\lambda_{T_{\max}}<C$ 时,则说明此步外循环已修复异常数据,内循环结束。

步骤5:假设在内循环结束后有 K 个异常数据在时刻 $T_1,T_2,\cdots,T_K$ 被识别出,其影响分别为 $\widetilde{\omega}_1^{(1)},\widetilde{\omega}_2^{(1)},\cdots,\widetilde{\omega}_K^{(1)}$,同时异常数据被修正而得到了新的时间序列 $\widetilde{Z}_t^{(1)}$(右上角的1表示这是第1次外循环迭代得到的序列)。此时重新回到步骤3,进入外循环,根据式(2)重新估计该时间序列参数 $\widetilde{\theta}^{(1)}(B)$,$\widetilde{\varphi}^{(1)}(B)$,$\widetilde{\pi}^{(1)}(B)$,并根据式(22)和式(23)得到时间序列模型残差 $\widetilde{e}_t^{(1)}$ 为:

$$\widetilde{e}_t^{(1)}=\widetilde{\pi}^{(1)}(B)\left(\widetilde{Z}_t^{(1)}-\sum_{j=1}^{K}\omega_j^{(1)}\widetilde{v}_j^{(1)}(B)I_t^{(T_j)}\right) \tag{28}$$

$$\widetilde{v}_j^{(1)}(B)=\begin{cases}\dfrac{\widetilde{\theta}^{(1)}(B)}{\widetilde{\varphi}^{(1)}(B)\nabla^d} & \mathrm{IO}\\ 1 & \mathrm{AO}\end{cases} \tag{29}$$

根据重估的时间序列参数计算检验统计量,当 $\lambda_{T_{\max}}<C$ 时外循环结束,当 $\lambda_{T_{\max}}>C$ 时重新进入外循环,直到所有的异常数据都被修复。

步骤6:在最后一次外循环结束后,针对修正了噪声点的时间序列 $\widetilde{Z}_t$ 进行联合估计,得到拟合异常值的模型。

$$\widetilde{Z}_t=\sum_{j=1}^{K}\widetilde{\omega}_j\widetilde{v}_j(B)I_t^{(T_j)}+\frac{\widetilde{\theta}(B)}{\widetilde{\varphi}(B)\nabla^d}a_t \tag{30}$$

式(30)中,各参数 $\widetilde{\theta}(B)$,$\widetilde{\varphi}(B)$,$\widetilde{\omega}_j$,$\widetilde{v}_j$ 是在最后一次迭代中得到的,该联合估计的目的是验证数据清洗的数学模型是否与真实数据相近,即拟合残差属于可接受范围。此时,将式(30)中异常时间点的数据作为"修正"的数据,以替代原始数据,而其他时间点的数据仍保留原始值。

步骤7:使用式(19)和式(21)的时间序列干预模型拟合原始数据,并求出干预点发生时间。

3　算例分析

3.1　数据清洗算例

算例1选取南方电网某输电线路采集的导线温度数据,如图1中实线所示,该时间序列不存在噪声点和缺失值,属于周期性时间序列,可用季节型ARIMA(p,d^s,q)拟合。为了检验本文数据清洗方法的实用性,将原来的观测时间点 $t=140$ 的数值剔除(成为缺失点),观测时间点 $t=26$ 和 $t=49$ 分别

加入一个 AO 和 IO 异常值，从而生成了一个带清洗的时间序列 Z_t，如图 1 虚线所示。

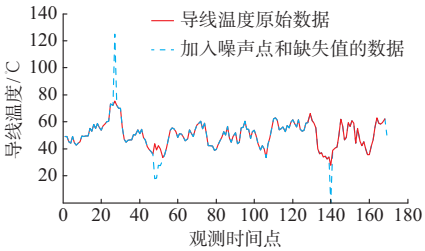

图 1　原始数据和含有异常值的数据
Fig.1　Original data and the simulated data with outliers

利用 MATLAB 软件对时间序列 Z_t 进行数据清洗，步骤如下。

步骤 1：进入外循环，首先对时间序列 Z_t 拟合季节型 ARIMA(p,d^s,q)，得到 ARIMA$(1,0,0)$ (s 为 12)如式(31)所示。

$$(1-0.914B)(1-B^{12})Z_t=1.02+a_t \quad (31)$$

观察拟合模型的残差序列初步推断该序列可能存在多个异常值。

步骤 2：进入内循环，计算每个观测点的检验统计量 T_{AO} 和 T_{IO}，逐次迭代直到所有的噪声点都被检验出来，结果如表 1 所示。

表 1　检验出的异常值类型
Table 1　Types of tested outliers

迭代次数	观测时间点	类型
1	140	AO
2	49	IO
3	26	AO

因此，考虑如下的修正模型：

$$Z_t=\theta_0+\omega_1 I_t^{(26)}+\omega_2\frac{1}{(1-\varphi_1 B)\nabla^{12}}I_t^{(49)}+\omega_3 I_t^{(140)}+\frac{1}{(1-\varphi_1 B)\nabla^{12}}a_t \quad (32)$$

式中：θ_0 为 $\theta(B)$ 的相应参数。

根据表中的拟合影响对时间序列的噪声点和缺失值数值进行修正，同时，根据式(32)对修正后的时间序列重新估计其参数，得到第 1 次修正后的时间序列及残差图如图 2 所示。

步骤 3：根据图 2 中的残差可判断原数据的噪声点和缺失值全部被检验了出来。但由于噪声点对于观测时间点的数值拟合残差过大，不符合赤池信息准则(AIC)检验，因此，需要返回外循环进行迭代计算，进一步修正时间序列，以提高数据清洗质量。

步骤 4：在通过两次外循环的迭代之后(逐步拟合结果见表 2)，得到最终清洗后的时间序列，如图 3 红色点所示，与原始数据基本符合。

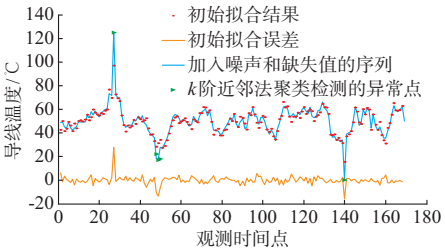

图 2　加入异常值的序列和初始拟合残差序列
Fig.2　Original series with outliers and the initial fitting residual series

表 2　逐步拟合结果
Table 2　Fitting results of each step

迭代次数	θ_0	φ_1	ω_1	ω_2	ω_3
1	1.093	0.868	0.821	−0.513	0.990
2	1.105	0.852	0.133	−0.320	0

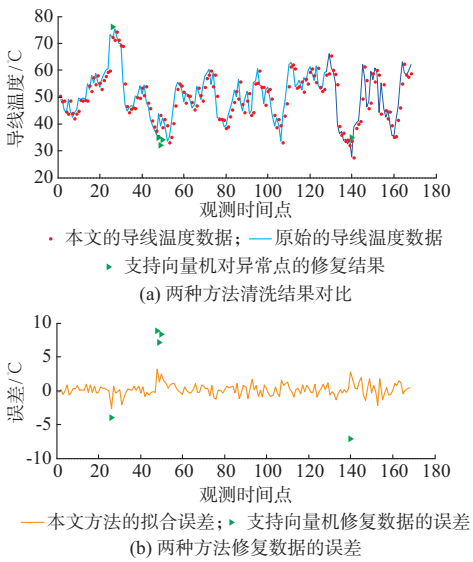

图 3　原始数据、最终拟合的数据、最终拟合误差和支持向量机的修复结果
Fig.3　Original data, final fitting data, the final fitting errors and the results of support vector machine

从图 3 可以看出，虽然清洗后的时间序列与原始时间序列在异常值发生时刻附近存在偏差，但是偏差都在 10% 以下，不影响后续的状态评估，属于可接受的范畴。作为对比，使用 k 阶近邻法聚类来检测数据中的噪声点和缺失值，并使用回归支持向量机修复数据。两种方法的对比结果如表 3 所示，由于温度数据不平稳且周期性变化的特点，支持向量机前向和后向预测结果差别大，精确度不如本文方法。图 3 中的绿色点表示检测出的噪声点，其结果与本文方法相近，检测出了时间点 26，48，49，50 为噪声点，时间点 140 为零值(缺失值)。

表 3　两种方法结果对比
Table 3　Results comparison of two different methods

时间点	原始数值/℃	本文方法修复值/℃	本文方法误差/%	支持向量机修复值	支持向量机误差/%
26	75.1	78.2	4.0	79.2	5.3
48	43.8	40.7	7.1	37.5	14.4
49	39.0	36.9	5.3	35.3	9.5
50	42.4	39.8	7.0	34.2	19.3
51	39.8	38.5	3.3		
140	27.9	28.9	3.6	33.9	21.5

算例 2 是对多元时间序列的数据清洗（见附录 B）。以上两个算例表明，基于时间序列的数据清洗方法是针对数据整体规律而言的，能够修复时间序列中的噪声点和缺失值，完成数据清洗的目标。支持向量机方法对局部平稳性或固定趋势性序列的清洗结果与本文方法相近，但是对非平稳或季节性数据清洗结果差，具有局限性。

3.2　干预模型算例

算例 3 为某变电站油中气体 CH_4 的在线监测数据，如图 5 所示，通过 ARIMA(1，0，1)模型拟合后得到数据的初始拟合结果和拟合误差，其中 CH_4 含量表示每升空气中 CH_4 的含量。可以看出，在观测时间点 $t=50$ 左右时间序列发生了趋势的改变，CH_4 气体的值由平稳趋势变为上升趋势，可以定性地判断变压器的内部绝缘出现劣化加速趋势。因此，针对此类异常数据，应使用时间序列的干预模型获取故障有效信息，不可用作数据清洗。

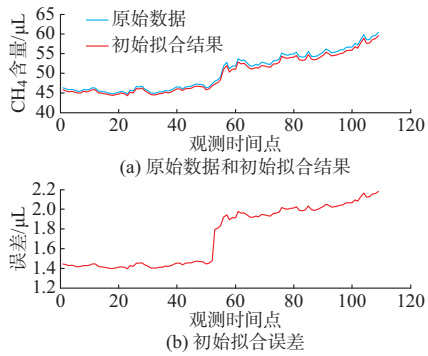

图 4　CH_4 气体的原始数据及初始拟合结果和误差
Fig.4　Original data of CH_4 and initial fitting results and errors

由于 $t=50$ 处气体数据发生了趋势改变，因此，可以用第 2 类干预结构来拟合原始数据：

$$(1-B)Z_t=\mu+\frac{\omega B}{1-\delta B}S_t^{(50)}+a_t \qquad (33)$$

异常数据的最终拟合结果如图 5 所示。从该干预模型可以得出结论，在 $t=50$ 处变压器出现了异常运行状态，需要运维人员密切关注。实际情况是变压器低压侧上夹件内衬铁斜边与C 相端处相碰，形成了故障接地点，从而与原接地点形成了环流，使得变压器过热，与本文方法结论一致。

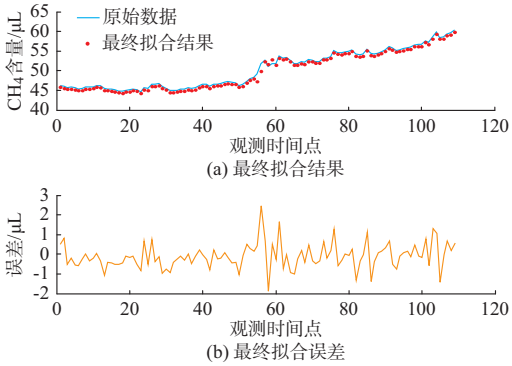

图 5　异常数据的最终拟合结果
Fig.5　Final fitting results of outliers

4　结语

本文基于时间序列分析这一大数据技术，利用模型拟合残差的迭代检验法来检验出输变电设备状态数据中的噪声点和缺失值，并能在迭代过程中对异常数据进行修正。相比于传统的删除噪声点，本文方法清洗出的数据是不带有噪声点和缺失值的数据，从而避免了时间序列中有用信息的丢失。3 个应用实例表明本文方法能实现输变电设备状态信息海量历史数据的校验和清洗，不仅可以自动去除“脏数据”，而且能够提升数据的质量，避免设备状态的误诊。

附录见本刊网络版（http://www.aeps-info.com/aeps/ch/index.aspx）。

参考文献

[1] 宁辽逸，吴文传，张伯明.运行风险评估中的变压器时变停运模型：(一)基于运行工况的变压器内部潜伏性故障的故障率估计方法[J].电力系统自动化，2010，34(15)：9-13.
NING Liaoyi，WU Wenchuan，ZHANG Boming. Time-varying transformer outage model for operational risk assessment：Part one　condition based failure rate estimation method for transformer internal latent fault estimation[J]. Automation of Electric Power Systems，2010，34(15)：9-13.

[2] 郭创新.采用分层多源信息融合的电网故障诊断方法[J].高电压技术，2010，36(12)：2976-2983.
GUO Chuangxin. Hierarchical fault diagnosis for power grid with information fusion using multi data resources[J]. High Voltage Engineering，2010，36(12)：2976-2983.

[3] 张金江，郭创新，曹一家，等.变电站设备状态监测系统及其 IEC 模型协调[J].电力系统自动化，2009，33(20)：67-72.
ZHANG Jinjiang，GUO Chuangxin，CAO Yijia，et al. Substation equipment condition monitoring system and IEC model coordination[J]. Automation of Electric Power Systems，2009，33(20)：67-72.

[4] 王慧芳，杨荷娟，何奔腾，等.输变电设备状态故障率模型改进分析[J].电力系统自动化，2011，35(16)：27-31.

WANG Huifang，YANG Hejuan，HE Benteng，et al. Improvement of state failure rate model for power transmission and transforming equipment[J]. Automation of Electric Power Systems，2011，35(16)：27-31.

[5] 王德文，邸剑，张长明.变电站状态监测 IED 的 IEC 61850 信息建模与实现[J].电力系统自动化，2012，36(3)：81-86.
WANG Dewen，DI Jian，ZHANG Changming. Information modelling and implementation for status monitoring IED in substation based on IEC 61850[J]. Automation of Electric Power Systems，2012，36(3)：81-86.

[6] 张海波，易文飞.基于异步迭代模式的电力系统分布式状态估计方法[J].电力系统自动化，2014，38(9)：125-131.
ZHANG Haibo，YI Wenfei. Distributed state estimation method for power systems based on asynchronous iteration mode[J]. Automation of Electric Power Systems，2014，38(9)：125-131.

[7] 张斌，张东来.电力系统稳态数据参数化压缩算法[J].中国电机工程学报，2011，31(1)：72-79.
ZHANG Bin，ZHANG Donglai. Parametric compression algorithm for power system steady data[J]. Proceedings of the CSEE，2011，31(1)：72-79.

[8] 宋亚齐，周国亮，朱永利.智能电网大数据处理技术现状与挑战[J].电网技术，2013，37(4)：927-935.
SONG Yaqi，ZHOU Guoliang，ZHU Yongli. Present status and challenges of big data processing in smart grid[J]. Power System Technology，2013，37(4)：927-935.

[9] CHEN Jiyi，LI Wenyuan，LAU A，et al. Automated load curve data cleansing in power systems[J]. IEEE Trans on Smart Grid，2010，1(2)：213-221.

[10] BRIGHENTI C，SANZ-BOBI M A. Auto-regressive processes explained by self-organized maps：application to the detection of abnormal behavior in industrial processes[J]. IEEE Trans on Neural Networks，2011，22(12)：2078-2090.

[11] MESSINA A R，VITTAL V. A structural time series approach to modeling dynamic trends in power system data[C]// Proceedings of 2012 IEEE Power and Energy Society General Meeting，July 22-26，2012，San Diego，USA：8p.

[12] 叶鸥，张璟，李军怀.中文数据清洗研究综述[J].计算机工程与应用，2012，48(14)：121-129.
YE Ou，ZHANG Jing，LI Junhuai. Survey of Chinese data cleaning[J]. Computer Engineering and Applications，2012，48(14)：121-129.

[13] 魏武雄.时间序列分析：单变量和多变量方法[M].北京：中国人民大学出版社，2009.

[14] 吴立增，朱永利，苑津莎.基于贝叶斯网络分类器的变压器综合故障诊断方法[J].电工技术学报，2005，20(4)：45-51.
WU Lizeng，ZHU Yongli，YUAN Jinsha. Novel method for transformer faults integrated diagnosis based on Bayesian network classifier[J]. Transactions of China Electrotechnical Society，2005，20(4)：45-51.

[15] YANG Xiaowei，ZHANG Guangquan，LU Jie，et al. A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises[J]. IEEE Trans on Fuzzy Systems，2011，19(1)：105-115.

[16] BATUWITA R，PALADE V. FSVM-CIL：fuzzy support vector machines for class imbalanced learning[J]. IEEE Trans on Fuzzy Systems，2010，18(3)：558-571.

[17] BRANDT P T，WILLIAMS J T. Multivariate time series model[M]. USA：SAGE Publications Inc.，2006.

[18] 王振龙.应用时间序列分析[M].北京：中国统计出版社，2010.

[19] IEEE Standard C57 104—2008 IEEE guide for the interpretation of gases generated in oil-immersed transformers[S]. 2009.

[20] 国家电网公司运维检修部.变压器类设备典型故障案例汇编[M].北京：中国电力出版社，2012.

严英杰(1988—)，男，通信作者，博士，主要研究方向：输变电设备状态评估。E-mail：yanyingjie@sjtu.edu.cn

盛戈皞(1974—)，男，教授，主要研究方向：输变电设备智能化技术。E-mail：shenghe@sjtu.edu.cn

陈玉峰(1970—)，男，高级工程师，主要研究方向：输变电设备管理及状态检修。

（编辑　万志超）

Cleaning Method for Big Data of Power Transmission and Transformation Equipment State Based on Time Sequence Analysis

YAN Yingjie[1]，*SHENG Gehao*[1]，*CHEN Yufeng*[2]，*JIANG Xiuchen*[1]，*GUO Zhihong*[2]，*QIN Shaopeng*[3]

(1. School of Electronic Information and Electrical Engineering，Shanghai Jiao Tong University，Shanghai 200240，China；
2. Electric Power Research Institute of State Grid Shandong Electric Power Company，Jinan 250002，China；
3. Guang'an Power Supply Company of State Grid Sichuan Electric Power Company，Guang'an 638500，China)

Abstract： Data cleaning is a key step in data preprocessing for state assessment of power equipment to help improve data quality and utilization. As the device status information can be made equivalent to the multivariate time sequence of each state，an iterative data cleaning method based on time sequence analysis is proposed. First，the abnormal data in time sequence is classified with the missing values treated as one of the types of the anomalies. Then the impact of different types of anomalies on the sequential model is quantified and several implementation steps of the iterative method are described. Finally，the approach is tested on the on-line monitoring data of a power equipment of the China Southern power grid. The results show that this method is capable of not only effectively identifying the abnormal data，but also repairing the noise points and missing values in meeting the data cleaning requirement.

This work is supported by National Natural Science Foundation of China (No. 51477100)，National High Technology Research and Development Program of China (863 Program) (No. SS2012AA050803) and State Grid Corporation of China.

Key words： big data；data cleaning；time sequence；state data of power equipment