用电采集系统停上电事件的数据质量辨识技术

王建雄[1],罗心仪[2],闫　林[2],唐海国[1],马　瑞[2]

(1.国网湘潭供电公司,湖南 湘潭 411100;2.长沙理工大学电气与信息工程学院,湖南 长沙 410114)

摘　要:为解决智能配电网抢修服务平台中数据质量缺陷对预警抢修精准度产生负面影响的问题,提出一种针对用电信息采集系统停上电信息的数据质量处理方法。首先使用回归方法对数据整体异常率进行辨识,其次依据完整性、唯一性、一致性、准确性各项辨识指标对停上电数据缺陷进行辨识与处理,最后分析最新数据与历史数据之间的显著关系,辅助实现数据时效性辨识。该方法对源数据的质量进行辨识与处理,为后期平台建设中信息集成和故障研判的开发运行提供可靠的数据支撑。

关　键　词:智能配网;数据质量;异常检测;停上电数据;用电采集系统

DOI:10.19781/j.issn.1673-9140.2021.05.023　　**中图分类号**:TM732　　**文章编号**:1673-9140(2021)05-0187-08

Identification technology of power-off event data quality in electricity acquisition system

WANG Jianxiong[1], LUO Xinyi[2], YAN Lin[2], TANG Haiguo[1], MA Rui[2]

(1. State Grid Xiangtan Power, Changsha 410007, China; 2. School of Electrical Engineering, Changsha University of Science and Technology, Changsha 410114, China)

Abstract: The data quality defects in the emergency repair service platform of intelligent distribution network have negative impacts on the accuracy of early warning and emergency repair. Under the background, a data quality processing method for power-off information of power consumption information acquisition system is proposed. Firstly, the regression method is utilized to identify the overall anomaly rate of the data. Secondly, the defects of the shutdown and power on data are identified and processed according to the identification indicators of integrity, uniqueness, consistency and accuracy. Finally, the significant relationship between the latest data and historical data is analyzed to help identify data timeliness. This method identifies and processes the quality of source data, and provides reliable data support for the development and operation of information integration and fault diagnosis in the later platform construction.

Key words: intelligent distribution network; data quality; anomaly detection; power-off data; electricity acquisition system

收稿日期:2018-09-01;**修回日期**:2019-11-08

基金项目:国家自然科学基金(51277015);国网湖南省电力公司科技项目(5216A5160031)

通信作者:王建雄(1973-),男,硕士,主要从事配电网自动化技术研究;Email:yanlin2412@163.com

随着大数据环境的出现，如果能有效地组织和使用大数据，将对社会经济和科学研究发展产生巨大的推动作用，同时也孕育着前所未有的机遇[1]。电力行业面对着新的成长机会，传统配电网由被动模式向主动模式逐渐转变，并提出配电网故障自动化智能抢修与主动服务模式相结合的新路线。另外，地区配电网在处理停上电事件的过程中，由于缺乏规范的规则库，通常因配电网运行的复杂性和相关数据的多样性，导致产生虚假信号。由于现有配电网抢修指挥平台的源头数据存在 30～40％的误报，严重影响了主动抢修的效率和研判准确性，导致生成大量的误报工单。因此，为顺应大数据环境带来的行业大数据研究热潮，必须解决存在的问题和面对现有的挑战。为了给配电网故障的智能抢修提供有效地数据支撑，提出针对配电网应用平台的数据质量评估方法，是必不可少的研究环节。

目前对电力数据质量的研究大多集中在电力系统安全性和电能质量方面，对配电服务所需数据的处理并没有一套规范的方法探讨。文献[2]提出了一种智能配电网多维数据质量评价方法，通过多维度分析和决策树对智能配电网数据多层面、多方位、多角度分析和挖掘；文献[3]提出必须有效地处理广域信息大数据，由此提出了建立数据可信分级、设计专用数据处理控制器、增添数据分群管理模式等数据应用策略；文献[4]提出多源多时空信息的综合检测方法及判断依据，能够有效地辨识和修正配电网 SCADA 系统中不满足精度要求的电压数据；文献[5]设计了适用于电力企业的相关数据治理体系，并分析了影响电力数据质量的主要因素，根据数据质量的一致性、准确性、完整性和及时性，建立了数据质量评估指标。

该文提出一种针对用电信息采集系统停上电信息的数据质量处理方法，深度融合传统大数据辨识法则、电力数据停上电信息特性以及配电网平台后续的主动配电及抢修需求，为停上电数据选择合适的质量辨识指标，研究了停上电数据质量的辨识逻辑，解决了数据质量缺陷对后续算法模型精准度产生负面影响的问题，为电力采集系统的应用提供可靠的数据支撑，使自动化智能服务平台更加精准化。

1 停上电事件数据质量评估

用户采集系统的停上电数据质量评估，实质上就是评判一个存在多种关联关系的设备推送停上电信号时，是否发生错推、漏推的现象，使得一个或多个设备由于源头数据的缺失和错乱，发生数据误报的现象。

首先，分析停上电事件数据的整体异常率；其次，选择合适的评估指标；然后，制定高效的辨识规则并判断异常存在的情况；最后，根据设备关联关系填充或修正这些问题数据，建立一个闭环的停上电事件数据质量评估模型。

1.1 停上电数据质量评估流程

辨识和纠正停上电事件数据缺陷的流程如图 1 所示。为避免配电网中明显有误数据的无效录入，必须从整体上对待测数据进行异常检测，剔除异常数据记录，提高停上电事件数据整体质量。针对抢修数据中常有缺项漏项的情况，必须保证不出现因数据异常缺失而导致的大量无效主动工单，进而影响后续抢修工作和数据处理，故设置完整性评估指标[6-8]。为保证数据的录入正确，便于整理和分析，设置一致性和准确性评估指标。针对数据冗余情况，且在抢修平台中既要保证事件编号唯一，不发生一号多事，同时也防止多号一事，即防止重复录入 2 次同一地点和时间的同样事件的数据以及防止误录造成的其他错误，比如整行、整列重复录入的错误，故设置唯一性评估指标。为保证事件的时效，设置

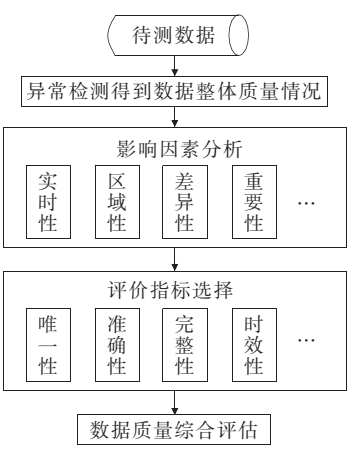

图 1 数据质量辨识流程

Figure 1 Data quality identification process

时效性评估指标[9-10]。根据配电网的数据特性设置与其匹配的评估指标，可以提高停上电事件数据质量辨识的准确性及可靠性。

2　停上电数据质量辨识的实现

对系统中的停上电事件数据进行多指标质量辨识前，需要对数据进行一次整体性的异常辨识，查找可能有异常的数据，确定其整体数据质量情况，即数据情况的总体异常比率。

2.1　停上电数据的异常辨识

考虑到停上电数据的多数变量之间不存在直接逻辑关联，故采用回归的方法对数据进行异常辨识[11]，用于检验数据的整体有效性。首先确定方程中的被解释变量与解释变量；其次确定回归模型；然后建立回归方程；最后对回归方程进行检验。

由于不确定回归方程是否能够精确的预测以及是否能够准确反映事物之间的统计关系，则需要进行回归方程的拟合优度检验。另外还要进行回归方程及回归系数的残差分析、显著性检验、多重共线性检验。

① 回归方程的拟合优度检验。常用于回归方程拟合优度的确定，可决系数指标为

$$\begin{cases} R^2=\dfrac{S_{SR}}{S_{ST}}=1-\dfrac{S_{SE}}{S_{ST}} \\ S_{ST}=\sum\limits_{i=1}^{n}(y_i-\bar{y})^2 \\ S_{SR}=\sum\limits_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\ S_{SE}=\sum\limits_{i=1}^{n}(y_i-\hat{y})^2 \end{cases} \tag{1}$$

式中　S_{ST} 为偏差的平方总和，用于反映被解释变量的总变化量；S_{SR} 为变差解释或回归平方和，可用回归方程来进行解释；S_{SE} 为剩余变异平方差或方差平方和，用于反映随机因素对被解释变量总变动的影响程度，属于回归方程不能解释的部分。

由于解释变量指标个数增加会使 R^2 增加，故采用用于调整的可决系数，其数学表述为

$$\bar{R}^2=1-\frac{S_{SE}/(n-k-1)}{S_{ST}/(n-1)} \tag{2}$$

② 回归方程的显著性检验。首先，建立 2 个原假设，假设 1 表示回归方程整体不显著，即 $\beta_1=\beta_2=\cdots=\beta_j=0$；假设 2 表示回归方程整体显著，即 β_j 不全等于 0。其中，$j=1,2,\cdots,k$。

对解释变量的方差进行分析，构建统计量为

$$\begin{cases} F=\dfrac{M_{SR}}{M_{SE}} \\ M_{SR}=\dfrac{S_{SR}}{k-1} \\ M_{SE}=\dfrac{S_{SE}}{n-k} \end{cases} \tag{3}$$

式中　$k-1$ 为解释变差的自由度；$n-k$ 为剩余高差的自由度；M_{SR} 为解释变差的均方差；M_{SE} 为剩余变差的均方差。

F 统计量及其对应的 p 值是使用零假设和样本数据计算的，其中 p 值表示检验假设中零假设成立或表现更严重的可能性。对比 p 和 α 值，结合原假设做出判断，其中 α 为选定显著性水平。如果 $p<\alpha$，则拒绝原假设，认为回归方程整体显著；如果 $p>\alpha$，接受零假设，即回归方程整体非显著。

③ 回归系数的显著性检验。首先，建立 2 个初始假设，初始假设 1 表示第 j 个回归系数非显著，即 $\beta_j=0$；初始假设 2 表示第 j 个回归系数显著，即 $\beta_j\neq0$。

其次，构造统计量为

$$t_j=\frac{\hat{\beta}_j}{\sqrt{\sum\limits_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}} \tag{4}$$

该统计量服从满足零假设条件下自由度 $(n-k-1)$ 的 t 分布，即

$$\hat{\delta}=\sqrt{\frac{\sum\limits_{i=1}^{n}(y_i-\hat{y})^2}{n-2}} \tag{5}$$

比较 p 和 α，并根据原始假设做出判断。如果 $p<\alpha$，原始假设被拒绝并且第 j 个回归系数被认为是显著的；如果 $p>\alpha$，则第 j 个回归系数非显著的零假设成立。

④ 残差分析。残差是指真实样本值与回归方程预测结果之间的差值，即

$$e_i=y_i-\hat{y}_i=y_i-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_ix_i) \tag{6}$$

经过残差分析过程，可以使随机扰动项对经典

假定的服从与否作出检验。

⑤ 多重共线性检验。对于解释变量 x_j 的容忍度为

$$T_{\text{ol-}j}=1-R_j^2 \tag{7}$$

式中 R_j^2 为方程中其他的解释变量与解释变量 x_j 之间的复相关系数平方,其作用是分析 x_j 与其他解释变量的线性相关性。显然,x_j 和其他解释变量的线性相关程度与 R_j^2 的大小成正相关,当 R_j^2 越大时,$T_{\text{ol-}j}$ 则越小。当 $T_{\text{ol-}j}$ 等于 0 时,变量 x_j 无多重共线性。

⑥ 探测样本中的异常值。在上述步骤 4 分析完残差后,对残差进行进一步处理和比对,从而实现样本异常值的探测,具体方法有以下 3 种[12-13],在该文中将会结合使用,发挥优化效果。

1)标准化残差。因为残差是均值为 0 的高斯分布,所以可以依据 3σ 准则进行判别。首先进行归一化残差,即

$$Z_{\text{RE-}i}=\frac{e_i}{\hat{\sigma}} \tag{8}$$

式中 $\hat{\sigma}$ 为回归方程的标准误差。

其次,观测 $Z_{\text{RE-}i}$ 的变化情况,观察值的绝对值如果大于 3,则该观察值是异常的。

2)学生化残差。由于普通残差分析不考虑异方差性,在异方差的情况下,应引入学生化残差的概念来判断其异常值,计算学生化残差为

$$S_{\text{RE-}i}=\frac{e_i}{\hat{\sigma}\sqrt{1-h_i}} \tag{9}$$

式中 h_i 为第 i 个样本的杠杆值。

其次,观测 $S_{\text{RE-}i}$ 的变化情况,如果观察值的绝对值大于 3,则该观察值是异常的。

3)剔除残差。剔除残差的逻辑是在总容量为 n 的停上电数据样本中,先排除要计算的第 i 个数据样本,再利用剩余数据样本重新拟合回归方程,并得出第 i 个样本的残差值与预测值。计算出的残差不仅与第 i 个样本无相关关系,而且第 i 个样本的 y 值是否异常也不会对残差产生影响。相较于残差检测,剔除残差的方式更能真实反映出第 i 个样本 y 值是否为异常的情况,是一种优化的残差分析思想。

由于在对配电网数据的处理中需要考虑异方差性,因此该文采用剔除残差的思想和学生化残差的概念相结合的异常检测方法,并规定了学生化残差被剔除后,如果观察值的绝对值大于 3,则确定该观察值为异常状态。

采用以上异常检测方法能确定数据的整体异常情况,找出明显错误,下一步将根据辨识指标对数据开展进一步异常辨识。

2.2 辨识指标的确立和数学实现

经过异常检测后的停上电数据,确定数据整体异常情况,该文后续将进一步分析其异常的具体原因,采用以下指标对数据进行进一步辨识。

1)完整性。数据完整性辨识流程如图 2 所示。考虑到用采系统数据间隐藏关系(标签)不明确的特性,采用 EM 算法对用采系统停上电数据完整性的进行检测[14]。

每个需要测试的样本都被认为是一个二元组 (x_i,z_i),其中,x_i 为第 i 个样本的观测值,z_i 为第 i 个样本的标签。

其中 E 步,对于每一个 i,其计算为

$$Q_i(z^{(i)})=p(z^{(i)}\mid x^{(i)};\theta) \tag{10}$$

M 步,计算为

$$\theta=\arg\max_{\theta}\sum_i\sum_{z^{(i)}}Q_i(z^{(i)})\log\frac{p(x^{(i)},z^{(i)},\theta)}{Q_i(z^{(i)})} \tag{11}$$

2)唯一性。配电网统计数据的是否唯一,需从以下 2 个方面进行分析:停上电事件数据记录中是否有完全相同的统计指标名称;不同横纵行列的数值数据是否完全一致,或相同数字的个数是否超过了某一规定允许阈值。

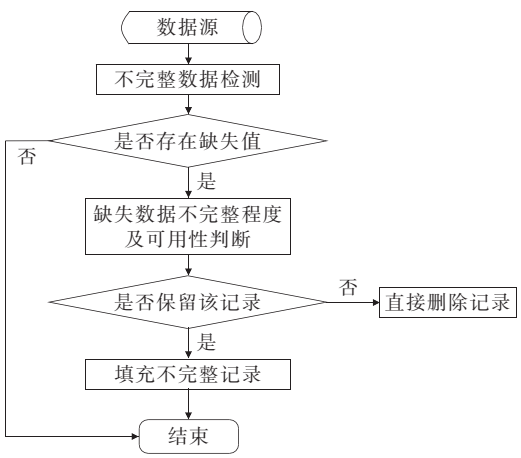

图 2 数据完整性辨识流程

Figure 2 Data integrity identification process

考虑以上条件，对配电网数据唯一性检测采用的逻辑为一旦统计表符合以上 2 点中的任何一点，则认为存在数据重复的问题。对于重复数据的初步检测，在最终确定是否为“真实”重复之前，还需要使用领域知识进行分析和判断。

3）一致性。一致性分析针对的是数据的表达格式，即要求同一属性下的数据使用相同的表达格式。在处理电气数据处理时，由于大多数数据都面对数值数据，并基本上以纯数字的形式描述，所以一致性分析可以降低为基于比率的数据。

而对于比率类数据，其有 3 种表达格式，即小数、“%”和“/”。这表明一致性可以通过以下原则进行分析：提前设置参考格式，然后通过扫描属性下的所有数据，将每种情况下的表达格式与参考格式进行比较，查看这两者是否一致，如果有差异，则认为数据不满足一致性的要求。

4）准确性。准确性分析是检查数据记录的精度是否符合要求。其检测原理与一致性类似，同样需要预先设定参考的精度值，然后再考察数据集中各数据记录的精度是否满足预先设定的参考精度值。

准确性检测逻辑如下，首先，将所有停上电事件记录看作字符串形式；然后通过计算“.”后的字符个数以得到个案的精度，而对于不存在“.”的个案，则直接将精度置 0；最后，再将个案的精度与参考精度进行对比，对于不满足参考精度的个案，则认为该个案不具备准确性。在准确性检测中，如有需要，还可对不满足准确性要求的数据进行相应的精度转换以使个案满足准确性要求。

5）时效性。停电信息由各部门分别管理，缺乏完整性、一致性和及时性[12]。数据的及时性即所谓的时效性，是指随着时间的推移和行业发展的日新月异，历史数据能否体现出最新数据的全部本质特征，并能对最新数据进行描述或替代，而不被历史所淘汰。其所衡量的是一种历史数据的可用性和有效性。对停上电数据的各项指标进行时效性检测，判断停电时间与上电时间的变化趋势是否有关联性，即综合判断停电事件是否得到有效的上电。

考虑到可能存在小规模的配电网统计样本，因此使用 t 检验来检验样本。其一般步骤如下。

第 1 步。建立虚无假设：$u_1=u_2$，即先假设 2 个总体平均数是不存在显著性差异的。

第 2 步。统计量 t 值的计算，针对不同的问题需选择相适应的计算方式。

若分析整体停上电数据样本平均值与小规模停上电数据样本平均数的差别程度，则计算统计量的数学形式为

$$t=\frac{\bar{X}-u_0}{\sqrt{\frac{S}{n-1}}} \tag{12}$$

式中　$\bar{X}$ 为样本的平均数；u_0 为总体的平均数；S 为样本的方差；n 为样本容量。

若分析 2 组样本的差别程度，则计算统计量值的数学形式为

$$t_2=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sum x_1^2+\sum x_2^2}{n_1+n_2-2}\times\frac{n_1+n_2}{n_1\times n_2}}} \tag{13}$$

式中　n_1、n_2 分别为样本 1、2 的容量值；x_1、x_2 分别为样本 1、2 的标准差；$\bar{X}_1$、$\bar{X}_2$ 分别为样本 1、2 的平均值。

第 3 步。根据自由度 $f=n-1$，比较 t 理论值，得到具体的自由度水准，判断数据间的差异性。0.01 级或 0.05 级可以设定为理论值差异的显著性水平，自由度不一样的显著性水平理论值记作 $t(f)0.01$ 和 $t(f)0.05$。

第 4 步。将理论 t 值与计算得出的 t 值进行比较，并且判断差异出现的概率，根据差异显著性关系以及 t 值的情况作出具体辨识，如表 1 所示。

表 1　t 值与差异显著关系

Table 1　t value and difference significant relationship

t 值	p 值	差异显著程度
$\geqslant t(f)0.01$	$\leqslant 0.01$	非常显著
$\geqslant t(f)0.05$	$\leqslant 0.05$	显著
$<t(f)0.05$	>0.05	不显著

3　算例分析

1）停上电数据异常辨识。对某配电网 2016 年 9 月 1 日，上报召测标志为 0，事件有效性标志为 1，

事件类型为 0(即上报的有效终端或电表停电事件)的数据进行整体异常情况辨识,异常辨识待测变量为 EVENT_TIME,INPUT_TIME,POWER_OFF_TIME,POWER_ON_TIME 4 项。回归分析结果如表 2、3 所示。

由于该组数据中 EVENT_TIME 和 POWER_OFF_TIME 所含数的数值应为相同,且检测结果亦确认此 2 组数据相同,因此在对其他数值进行辨识时,排除 POWER_OFF_TIME,仅用 EVENT_TIME 作为参考变量。

对 INPUT_TIME 回归计算结果是:标准残差大于 3 的异常数值共 146 个,占该项总比 0.6%。

对 POWER_ON_TIME 回归计算结果为标准残差大于 3 的异常数值 1 个。

2)停上电数据完整性。该文选择 EM(expectation-maximization algorithm)算法来完成缺失值的检测。

对某电网 2016 年 9 月 1 日,上报召测标志为 0,事件有效性标志为 1,事件类型为 0(即上报的有效终端或电表停电事件)的数据进行完整性检测,得到的结果如表 4 所示。

由表格统计结果可见在当日所收集的有效终端或电表停电事件中,共缺少 17 816 个上电时间数据,占应有上电时间数据的 76.3%。

3)停上电数据唯一性、一致性、准确性。考虑到停上电事件大多为时间数据,因此唯一性评估时,只考虑事件编号的唯一性。即每一个事件的录入必须有唯一编号对应,若出现重复或者缺失,则认为数据录入无效。

在对某电网 2016 年 9 月 1 日,上报召测标志为 0,事件有效性标志为 1,事件类型为 0(即上报的有效终端或电表停电事件)的数据进行唯一性、一致性、准确性检测的过程中,并未发现不符合检测标准的数据。

表 2 因变量为 INPUT_TIME 的回归分析

Table 2 Regression analysis with dependent variable INPUT_TIME

模型	非标准化系数		标准化系数	T	显著性	B 的 95%置信区间		共线性统计	
	B	标准误差	Beta			下限值	上限值	容许	VIF
1(Constant)	340 567.072	40 610.446	—	8.386	0.000	260 954.629	420 179.516	—	—
POWER_ON_TIME	0.003	0.005	0.007	0.496	0.620	−0.008	0.013	1.000	1.000
EVENT_TIME	−6.994	0.953	−0.098	−7.339	0.000	−8.863	−5.126	1.000	1.000

表 3 因变量为 POWER_ON_TIME 的回归分析

Table 3 Regression analysis with dependent variable POWER_ON_TIME

模型	非标准化系数		标准化系数	T	显著性	B 的 95%置信区间		共线性统计	
	B	标准误差	Beta			下限值	上限值	容许	VIF
1(Constant)	−57 817.722	104 802.024	—	−0.552	0.581	−263 270.898	147 635.454	—	—
EVENT_TIME	2.340	2.456	0.013	0.953	0.341	−2.474	7.154	0.990	1.010
INPUT_TIME	0.017	0.034	0.007	0.496	0.620	−0.050	0.085	0.990	1.010

表 4 停上电数据完整性检测结果统计

Table 4 Power-off data integrity test result statistics

模型	样本总数	平均值	标准偏差	缺失		极端数目	
				计数	百分比	低	高
EVENT_ID	23 347	410 008 233 474 695.50	293 746 652.2	0	0	0	2 274
EVENT_TIME	23 347	42 614.446 930 143 135	0.225 352 546 5	0	0	152	949
INPUT_TIME	23 347	42 621.099 577 897 294	18.272 660 18	0	0	0	1 266
POWER_OFF_TIME	23 347	42 614.466 930 143 135	0.225 352 546 5	0	0	152	949
POWER_ON_TIME	5 531	42 614.098 063 942 736	36.576 250 82	17 816	76.3	1	0
CP_NO	23 347	—	—	0	0	—	—
ORG_NO	23 347	—	—	0	0	—	—

4）停上电数据时效性。时效性指将历史数据与最新数据进行显著性分析，借以判断两者之间是否存在显著性差别。对停上电数据，时效性检验的作用是判断停电时间与上电时间的变化趋势是否有关联性即综合判断停电事件是否得到有效的上电。

在对某电网 2016 年 9 月 1 日，上报召测标志为 0，事件有效性标志为 1，事件类型为 0（即上报的有效终端或电表停电事件）的数据进行时效性检测，所得结果如表 5～7 所示。

表 5　配对样本相关性

Table 5　Paired sample correlation table

配对比较	数字	相关系数度	显著性
配对 1 EVENT_TIME INPUT_TIME	23 347	−0.060	0.000
配对 3 EVENT_TIME POWER_ON_TIME	5 531	0.012	0.364

表 6　配对样本统计

Table 6　Paired sample statistics

配对比较	平均值	数字	标准偏差	标准误差平均值
配对 1 EVENT_TIME	42 614.466 930 141 95	23 347	0.225 352 547 835 9	0.001 474 846 870 2
INPUT_TIME	42 621.099 577 896 72	23 347	18.272 660 177 481 8	0.119 587 623 624 8
配对 2 EVENT_TIME	42 614.466 930 141 95	23 347	0.225 352 578 359 0	0.001 474 846 870 2
POWER_OFF_TIME	42 614.466 930 141 95	23 347	0.225 352 578 359 0	0.001 474 846 870 2
配对 3 EVENT_TIME	42 614.444 882 631 79	5 531	0.201 281 215 378 9	0.002 706 458 778 9
POWER_ON_TIME	42 614.098 063 942 85	5 531	36.576 250 816 722 4	0.491 810 002 922 3

表 7　配对样本检测

Table 7　Paired sample test

配对比较	配对差值					T	自由度	显著性（双尾）
	平均值	标准偏差	标准误差平均值	差值的 95%置信区间				
				下限	上限			
配对 1 EVENT_TIME INPUT_TIME	−6.632 647 75	18.287 644 38	0.119 685 689 5	−6.867 239 56	−6.398 055 95	−55.417	23 346	0.000
配对 3 EVENT_TIME POWER_ON_TIME	0.346 818 688 9	36.574 345 57	0.491 784 384 6	−0.617 272 005	1.310 909 383	0.705	5 530	0.481

由检测结果和表 1 的 P 值与显著性关系可知，对于 EVENT_TIME 和 POWER_OFF_TIME，标准误差平均值相等，认为这 2 列数据具有一致性，不进行显著性评估；对于 EVENT_TIME 和 INPUT_TIME 的方差显著性和双尾显著性都是 0.00，说明 2 组数据方差和均值均不相等，且差值的 95%置信区间没有跨 0，说明这 2 组数据存在明显的差异；对于 EVENT_TIME 和 POWER_ON_TIME 的方差显著性大于 0.05，双尾显著性亦大于 0.05，说明 2 组数据不能拒绝方差相等假设和均值相等假设，并且差值的 95%置信区间跨 0，说明 2 组数据不存在明显的差异，且具有显著性。

4　结语

该文针对自动化配电网调度平台运行过程中所接收到的成千上万条停上电数据，提出一种数据质量辨识方法。为停上电数据选择合适的辨识指标进行数据质量的评估，找出停上电数据缺失项，过滤重复冗余数据，去除数据中的干扰噪声，并分析了最新数据和历史数据之间的显著关系，从而帮助实现对数据时效性的辨识。

通过该文提出的方法流程，可以得到配电网停上电数据的整体质量情况，找到异常点，为后期平台建设中信息集成和故障研判的开发运行提供了可靠的数据支撑，对于配电网自动化平台的建设具有一定的实用价值和学术价值。

基于这一技术，还需要考虑数据技术和配电网自动化平台的深度整合。如何结合其他数据对停上电数据进行进一步处理，如何保证数据清洗的持续性，达到数据清洗流程的闭环要求。参考停上电数据清洗技术，发展配电网平台中其他数据的清洗技

术,为配电网平台的发展及整个电力行业的研究提供优质准确的数据将成为需要继续研究深入考虑的问题。进一步深化数据应用,提高业务管理能力,为建设愈发完善的智能配电网抢修应用平台提供高质量的数据支撑。

参考文献:

[1] 李国杰,程学旗. 大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊,2012,27(6):647-657.
LI Guojie,CHENG Xueqi. Research status and scientific thinking of big data[J]. Bulletin of Chinese Academy of Sciences,2012,27(6):647-657.

[2] 潘旭,王金丽,赵晓龙,等. 智能配电网多维数据质量评价方法[J]. 中国电机工程学报,2018,38(5):1375-1384.
PAN Xu,WANG Jinli,ZHAO Xiaolong,et al. Multi dimensional data quality evaluation method for intelligent distribution network[J]. Proceedings of the CSEE,2018,38(5):1375-1384.

[3] 宋墩文,温渤婴,杨学涛,等. 广域量测信息大数据特征分析及应用策略[J]. 电网技术,2017,41(1):157-163.
SONG Dunwen,WEN Boying,YANG Xuetao,et al. Big data feature analysis and application strategy of wide area measurement information[J]. Power System Technology,2017,41(1):157-163.

[4] 刘科研,张剑,陶顺,等. 基于多源多时空信息的配电网SCADA系统电压数据质量检测与评估方法[J]. 电网技术,2015,39(11):3169-3175.
LIU Keyan,ZHANG Jian,TAO Shun,et al. Detection and evaluation of SCADA voltage data quality in distribution network based on multi temporal and spatial information of multi data sources[J]. Power System Technology,2015,39(11):3169-3175.

[5] 赵舫,朱彬若,王新刚. 基于短距无线的低压用户停电事件突发上报拥塞规避机制研究[J]. 电测与仪表,2020,57(5):125-131.
ZHAO Fang,ZHU Binruo,WANG Xingang. Research on congestion avoidance mechanism for burst report of power failure based on short distance wireless communications[J]. Electrical Measurement & Instrumentation,2020,57(5):125-131.

[6] 郭志民,马建伟,张小斐,等. 面向三维可视化场景的电力大数据分析模型构建研究[J]. 电网与清洁能源,2019,35(6):46-51.
GUO Zhimin,MA Jianwei,ZHANG Xiaofei,et al. Research on the construction of power big data analysis model oriented to 3D visualization scene[J]. Power System and Clean Energy,2019,35(6):46-51.

[7] 刘杰荣,张耀宇,关家华,等. 基于量测大数据和数学形态学的配电网故障检测及定位方法研究[J]. 智慧电力,2020,48(1):97-104.
LIU Jierong,ZHANG Yaoyu,GUAN Jiahua,et al. Distribution network fault location and detection method based on measurement big data and mathematical morphology[J]. Smart Power,2020,48(1):97-104.

[8] 白浩,袁智勇,梁朔,等. 基于大数据处理的配网运行效率关联性分析[J]. 电力系统保护与控制,2020,48(6):61-67.
BAI Hao,YUAN Zhiyong,LIANG Shuo,et al. Correlation analysis of distribution network operation efficiency based on big data processing[J]. Power System Protection and Control,2020,48(6):61-67.

[9] 和敬涵,刘琳. 基于冗余信息的智能变电站协同后备保护及故障预判算法研究[J]. 电力科学与技术学报,2015,30(2):3-8+22.
HE Jinghan,LIU Lin. Study of coordinated backup protection and fault forecast algorithm based on redundant information for smart substations[J]. Journal of Electric Power Science and Technology,2015,30(2):3-8+22.

[10] Markou M,Singh S. A neural network-based novelty detector for image sequence analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2006,28(10):1664-1677.

[11] 赵晋泉,张强,方嵩,等. 最小化用户停电损失的主动配电网黑启动分区优化策略[J]. 中国电力,2020,53(7):114-121.
ZHAO Jinquan,ZHANG Qiang,FANG Song,et al. Optimization strategy for black-start partitioning of active distribution network to minimize customer outage cost[J]. Electric Power,2020,53(7):114-121.

[12] 周剑,罗添允,李智勇,等. 基于改进可拓层次分析的停电影响综合评估[J]. 电力系统保护与控制,2019,47(3):31-38.
ZHOU Jian,LUO Tianyun,LI Zhiyong,et al. Comprehensive evaluation of power failure based on improved extension analytic hierarchy process[J]. Power System Protection and Control,2019,47(3):31-38.