DOI: 10.7667/PSPC161842

基于负荷预测与关联规则修正的不良数据辨识方法

王海云[1]，田莎莎[2]，张再驰[1]，陈 茜[1]，卢志刚[3]

(1. 国网北京市电力公司，北京 100000；2. 国网沧州供电公司，河北 沧州 061000；
3. 燕山大学电力电子节能及传动控制河北省重点实验室，河北 秦皇岛 066004)

摘要：随着电力系统的快速发展，使得电网需要对海量、异构和多态的数据进行分析与辨识。传统的不良数据辨识方法辨识效率较低，且不能够高效率利用已知的全部数据信息。为解决此问题，提出了一种基于负荷预测与关联规则修正的不良数据辨识方法。根据数据量之间的内在联系，给出了一种三维矩阵的数据存储方法。建立基于回归分析法的预测模型与基于灰色关联的相关性分析模型，分析节点注入功率与温度之间的变化关系，并采用关联规则与特殊断面修正法对预测值进行修正，进而完成对注入功率的辨识。在此基础上，再通过基尔霍夫定律与残差辨识法完成对支路潮流数据的辨识工作。最后应用实际系统的仿真算例证明了该方法能够在克服残差污染和残差淹没现象的前提下准确辨识出全部的不良数据。

关键词：不良数据辨识；数据存储；回归分析预测模型；相关性分析建模；关联规则

A new bad data identification method based on load forecasting and the correction of association rule

WANG Haiyun[1], TIAN Shasha[2], ZHANG Zaichi[1], CHEN Xi[1], LU Zhigang[3]

(1. State Grid Beijing Electric Power Company, Beijing 100000, China; 2. State Grid Cangzhou Power Supply Company, Cangzhou 061000, China; 3. Hebei Key Lab of Power Electronics for Energy Conservation and Motor Drive, Yanshan University, Qinhuangdao 066004, China)

Abstract: With the rapid development of power system, it is necessary for the grid to analyze and identify the massive, heterogeneous, polymorphic data. The efficiency of the traditional identification method is low, and it is not able to use all known data information efficiently. In order to solve this problem, a new bad data identification method based on load forecasting and the correction of association rule is proposed. At first, a three dimensional matrix data storage method is put forward according to the intrinsic relationship between the data. The forecast model of the regression analysis and the correlation analysis model based on grey correlation are established to analyze the relationship between the node power injection and the temperature. The association rules and special profile correction method are used to modify the predicted value, and then the identification of the injection power is completed. On this basis, the identification of the branch power flower is completed according to the Kirchhoff's law and the residual error identification method. At last, the simulation results of the practical system prove that the method proposed can overcome the “residual pollute” and “residual submerge” phenomenon to identify all the bad data accurately.

This work is supported by National Natural Science Foundation of China (No. 61374098) and Research Fund for the Doctoral Program of Higher Education of China (No. 20131333110017).

Key words: bad data identification; data storage; forecast model of the regression analysis; correlation analysis model; association rules

0 引言

电力系统规模与结构的日益复杂化以及区域电网的不断互联使得电力数据资源急剧增长，这样的电力系统将需要对整个系统的状态信息数据以及电网运行、检修和管理过程中产生的大量异构、多态的数据进行分析，进而能够迅速、准确并且全面地掌握电网的实际运行状态。因此，能否在大量数据中准确地辨识出不良数据成为了解决问题的关键步骤。

对生数据进行不良数据检测与辨识，并改正或

基金项目：国家自然科学基金(61374098)；教育部高等学校博士学科点专项科研基金(20131333110017)

删除不良数据，从而提高实时数据库的数据准确性、可靠性和完整性，是状态估计的基本功能之一[1-4]。

一个好的不良数据辨识方法在保证快速准确辨识的同时，还必须有较强的克服残差污染和残差淹没的能力。不良数据的检测与辨识是保证状态估计计算结果正确性的关键，因此一直是国内外学者们比较热衷的研究课题。目前，已经提出了多种不良数据辨识的方法，如残差搜索法(R_W 法或 R_n 法)、非二次准则法、零残差法及估计辨识法等。而用于辨识不良数据的新方法主要有基于数据挖掘的模糊数学法、神经网络法、聚类分析法及间隙统计法等。文献[5]提出一种新的多不良数据检测方法，利用拓扑方法推到了量测量之间的线性相关关系，并给出相应线性表达式。通过相关量测量之间的直接校核，提高了使检测相关多不良数据的可靠程度。文献[6]将标准化残差 R_n 检测法与状态预估检测辨识法相结合进行不良数据的检测与辨识。该方法用 R_n 检测法进行定性分析，即判断是否存在不良数据测点，再用状态预估辨识进行定位分析，并用预估值将不良数据进行替换。文献[7]提出了一种基于等效电流量测变换的计及不良数据检测与辨识的状态估计方法。该算法在无任何假设的情况下，能够实现雅可比矩阵的完全常数化。文献[8]提出了一种基于改进GSA 算法的电网不良数据检测与辨识，该算法从一定程度上避免了残差污染和残差淹没的现象，但计算量大。文献[9]提出一种基于遗传算法的改进迭代自组织分析法的电力系统不良数据辨识，该方法能最终获得全局最优解，并且避免了初始聚类中心对分类结果的影响。文献[10]分析了量测数据协方差矩阵中的元素值在量测数据中含有白噪声、突变量和不良数据时的变化规律，在此基础上，提出了一种通过协方差矩阵中元素的变化规律对不良数据进行检测和辨识的新方法。

虽然学者已经给出多种辨识不良数据的方法，但多数方法更适用于数据量不大的不良数据辨识问题，并且很难做到在保证辨识速度与精度的同时克服残差污染与残差淹没现象。因此，针对此问题，本文将负荷预测[11-16]方法中的回归分析法与相关性建模[17]以及关联规则挖掘[18-19]相结合，提出了一种能够在克服残差污染和残差淹没现象的前提下准确辨识出全部的不良数据。

本文首先根据数据量之间的内在联系，给出了一种三维矩阵的数据存储方法，建立基于回归分析法的预测模型与基于灰色关联的相关性分析模型，为分析节点注入功率与温度之间的变化关系打下基础。其次采用关联规则与特殊断面修正法对预测值进行修正，进而完成对注入功率的辨识，在此辨识结果的基础上，再结合基尔霍夫定律与残差辨识法完成对支路潮流数据的辨识工作。最后应用我国某一中等城市的实际输电网为算例的仿真结果证明了该方法能够在克服残差污染和残差淹没现象的前提下准确辨识出全部的不良数据。

1　注入功率的预测模型

根据电力系统负荷的特点，本文采用的是与频域分解相结合的回归分析法，首先通过频域分解，将总负荷分为日周期负荷分量、周周期负荷分量、低频负荷分量以及随机负荷分量，然后采用不同的方法针对不同性质的分量分别进行预测，最后将各个分量的预测值进行加和，得到最终的总负荷预测值。

1.1 总负荷在频域上的分解

电力系统中的总负荷是具有较强的周期性的时间序列，如白天与晚上的负荷特点一般不同，而冬季和夏季的负荷特点也不尽相同，这都体现着负荷的时间特性，因此能够将负荷分解为不同成分的分量，其分解方法可以采用频域分析方法。对于任意的负荷序列 $P(t)$其傅里叶分解结果为

$$P(t)=a_0+\sum_{i=1}^{N-1}(a_i\cos w_it+b_i\sin w_it) \tag{1}$$

式中，$w_i=\dfrac{i}{N}\times 2\pi, i=0,1,2,\cdots,N-1$。

其最终分解结果可表示为

$$P(t)=a+D(t)+W(t)+L(t)+H(t) \tag{2}$$

式中：$a_0+D(t)$ 为日周期负荷分量；$W(t)$ 为周周期负荷分量；$L(t)$ 为低频负荷分量；$H(t)$ 为高频负荷分量。

本文采用 $\mathrm{mod}(m,n)$ 表示 m 与 n 相除的余数，则上述四种分量对应角频率集合都可用与余数相关的集合表示。

$$\Omega_{\mathrm{day}}=\{w_0\}\cup\left\{w_i\left|\mathrm{mod}(24,\frac{2\pi}{w_i})\neq 0\right.\right\} \tag{3}$$

式中，Ω_{day} 为日周期负荷分量的角频率集合。

$$\Omega_{\mathrm{week}}=\left\{w_i\left|\mathrm{mod}(5\times 24,\frac{2\pi}{w_i})=0\text{且}\,\mathrm{mod}(24,\frac{2\pi}{w_i})\neq 0\right.\right\} \tag{4}$$

式中，Ω_{week} 为周周期负荷分量的角频率集合。

非周期性分量可分为低频分量 $L(t)$ 和高频分量 $H(t)$。低频分量 $L(t)$ 反映的是与气象因素相关的诸多因素对负荷的影响；高频分量 $H(t)$ 主要反映的是负荷变化的随机性，其角频率集合分别为

$$\Omega_{\text{low}}=\left\{w_i\left|\frac{2\pi}{w_i}\geq 24\text{且}w_i\notin\Omega_{\text{day}}\cup\Omega_{\text{week}}\right.\right\}\tag{5}$$

式中，Ω_{low} 为低频负荷分量的角频率集合。

$$\Omega_{\text{high}}=\left\{w_i\left|\frac{2\pi}{w_i}<24\text{且}w_i\notin\Omega_{\text{day}}\cup\Omega_{\text{week}}\right.\right\}\tag{6}$$

式中，Ω_{high} 为低频负荷分量的角频率集合。

通过上述过程，总负荷被分解成了两种周期变量和两种非周期变量。

1.2 回归分析预测法概述

基于回归分析的短期负荷预测方法的主要原理：首先对历史数据的变化规律与影响负荷变化的主要因素进行分析，假设负荷与一个或多个独立变量存在一定的因果关系，使用回归分析方法作为负荷预测的模型，采用最小二乘法对变量的观测数据进行分析，建立能够反映因果关系的数学模型来预测未来的负荷值，得出回归方程式，确定模型参数，根据此模型推断将来某个时刻的负荷值。回归分析法的外推性能好，对于历史上没有出现过的情况也能够有较好的预测。以下给出的是回归模型的一般形式。

假设变量 $x_1,x_2,\cdots,x_{\text{m}}$ 均与随机变量 y 之间存在一定的影响关系，一般情况下认为当 $x_1,x_2,\cdots,x_{\text{m}}$ 取定一系列值后，随机变量 y 的概率分布也随之而确定，并具有一一对应的特点。其模型可表示为

$$y=f(x_1,x_2,\cdots,x_{\text{m}})+\varepsilon\tag{7}$$

式中：y 为因变量；$x_1,x_2,\cdots,x_{\text{m}}$ 为自变量；$f(x_1,x_2,\cdots,x_{\text{m}})$ 为某确定性关系；ε 为随机误差。

本文进行节点注入功率短期预测时主要考虑的有：温度、湿度、降水量、节假日共四个影响因素，因此采用的模型为四元线性回归模型。

2　气象因素与节点注入功率间的相关性建模

不同的温度，湿度，降水量等气象因素对负荷会造成不同的影响，在电网结构中，节点所连接的负荷并不是同类型的，可以是第一产业、第二产业、第三产业、农村用电或者城镇用电，甚至是几种类型的混合负荷。但气象因素的变化只会引起部分负荷的变动，如第三产业中的商业住宿、餐饮业，以及居民用电等负荷。因此并不是所有节点的注入功率都对气象因素的变化敏感。在此环节中，对于相关性小于阈值的认为不相关，在挖掘关联规则时不考虑此类节点；对于大于阈值的，对其划分等级，有助于关联规则的挖掘。

进行相关性建模的方法有很多，如皮尔森系数法、灰色关联分析法[17]等。本文采用灰色关联分析法，其基本原理是依据曲线间的近似程度来判断因素间的关联程度。挖掘系统中各因素间的主要相互关系，寻找影响目标指标最严重的因素。本文建立的是节点注入功率与气象因素的相关性。

设本文要分析的目标指标即某节点的注入功率的序列为 $x_0=(x_0(1),x_0(2),\cdots,x_0(n))$；$x_0(1),x_0(2),\cdots,x_0(n)$ 表示不同时刻该节点的注入功率数值。系统行为序列，即影响注入功率的因素构成的序列为 $x_i=(x_i(1),x_i(2),\cdots,x_i(n)),i=1,2,\cdots,m$；$x_i(1),x_i(2),\cdots,x_i(n)$ 表示的是不同时刻该影响因素的数值。而不同的 i 代表的是不同的影响因素。

$$r(x_0(k),x_i(k))=\frac{\min\limits_i\min\limits_k|x_0(k)-x_i(k)|+\rho\max\limits_i\max\limits_k|x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)|+\rho\max\limits_i\max\limits_k|x_0(k)-x_i(k)|}\tag{8}$$

式中：$r(x_0(k),x_i(k))$ 为 x_i 对于 x_0 在 k 时刻的相关度；ρ 为分辨系数，一般取 $\rho=0.5$。

$$r(x_0,x_i)=\frac{1}{n}\sum_{k=1}^{n}r(x_0(k),x_i(k))\tag{9}$$

式中，$r(x_0,x_i)$ 为 x_i 对于 x_0 的相关度。

3　节点注入功率预测值的修正

3.1 基于关联规则的修正

关联规则是表示数据库中不同域之间具有满足某种指定要求的关联关系的规则。设 $I=\{i_1,i_2,\cdots,i_m\}$ 为项的集合，数据库 D 为与任务相关的事务数据库的集合，其中每个事务 T 表示为项的集合，$T\subseteq I$。每个事务都有一个标识符，可表示为：TID。假设 A 是一个相集，若项集 $A\subset D$，$B\subset D$，且 $A\cap B=\varnothing$，则形如 $A\Rightarrow B$ 的关系被称为关联规则。每个关联规则都具有两个特性参数，即支持度与置信度。支持度与置信度分别反映此关联规则的有效性和确定性，即支持度越高，说明其关联程度越高；置信度越高，可信度越高。其支持度与置信度的求解方法见式(10)与式(11)[18]。

$$\sup(A\Rightarrow B)=P(A\cup B)\tag{10}$$

式中：$\sup(A\Rightarrow B)$ 为某关联规则的支持度；$P(A\cup B)$ 为在全部事件中 A 与 B 同时发生的概率。

将 $\sup(A\Rightarrow B)\min$ 记为该关联规则的最小支持度阈值，一般取 70%，当支持度 $\sup(A\Rightarrow B)$ 大于 $\sup(A\Rightarrow B)\min$ 时，项集 A 为频繁项集。最小支持度阈值的取值可以根据研究情况的不同而适当调整，并非一成不变的。

$$\text{conf}(A\Rightarrow B)=P(B|A)\tag{11}$$

式中：$\mathrm{conf}(A \Rightarrow B)$ 为某关联规则的置信度；$P(B|A)$ 为在发生 A 事件时，同时发生 B 的概率。

支持度与置信度的求解过程实质就是求解两个概率的过程，只是在求解过程中，整体事件的选取不同。

在本节中，以温度为例进行关联规则挖掘的说明。选取完整的历史数据样本，分析其关联性，将关联规则记为：$\Delta T \Rightarrow \Delta P$。$\Delta T$ 表示的是温度的变化幅度，结合实际清况，本文的 $\sup(A \Rightarrow B)\min$ 选 10%。因此在本文中温度的变化幅度只需要研究 $\Delta T = 1, \Delta T = 2, \Delta T = 3$ 以及 $\Delta T = 4$ 的四种情况。

在进行关联规则的挖掘之前需要对历史数据进行处理，以为不同的运行方式或者日期类型都可能对关联规则的结果产生一定的影响。所以首先将每个数据断面按照(运行方式，日期类型)进行标记，如：运行方式若有 3 种，则标记出 1，2，3。日期类型指的是工作日与公休日，工作日为 1，公休日为 0。如：某断面被标记为(1,0)，则表示该断面采集当天为公休日，运行方式为第一种。然后按照此标记结果将所有断面进行分类，具有相同运行方式及日期类型的是一类。假设划分出了 k 类不同断面，进而对每一类断面挖掘其相应的关联规则。

以工作日为例对关联规则的挖掘方法进行说明。在此类中将所有断面进一步进行标记(日期，时刻，温度)。所谓的日期时刻指的是该断面采集当天的日期与时间，标记方式如下：1—31 分别对应某个月份的 1—31 日；将时刻数据折算成整数表示，如该断面的采集时间为 00:15，则直接表示为 15。若采集时间为 02:30，则表示为 $2 \times 60 + 30 = 150$，即表示为 150，依次类推。假设该类中有 n 个断面，将任意两个不同天但同时刻的断面的特征向量做差，分别找到两个项集的对应关系：[温差]—[相关节点的注入功率变化百分比]。在此过程中之所以要选择相同时刻，主要是为了避免由于不同时刻造成的功率差。此关联规则是随时更新的，以 30 天为一个时间窗，每 15 min 进行一次数据采集，每新测得一天的数据就需要更新一次关联关系，这样才能保证关联关系的实时性。

在得到 $\Delta T \Rightarrow \Delta P$ 此关联规则后，可运用此关联规则对节点注入功率的预测值进行验证，并将不满足该关联规则的数据进行修正，从而得到更为精确的注入功率预测值。

3.2 基于特殊断面的修正

当待计算断面当天发生了某特殊事件时，上述关联规则可能不再适用。本文考虑的特殊事件包括两类：节假日、重大会议或者大型比赛。针对这两种情况分别给出不同的数据修正方案。

3.2.1 针对节假日断面的修正

由于节假日负荷数据与正常日负荷数据相差较大，且节假日负荷数据相对较少，因此无法对节假日时负荷数据和气象因素之间的关联规则进行分析，即不能够挖掘出该节点在节假日时的注入功率与气象因素的关系。因此在本文中，采用差值分析法对节假日的注入功率进行估算。由于各年的节假日负荷具有很强的相似性，因此，可以通过参照前几年节假日注入功率曲线形状的特点，再结合本年负荷的实际水平，通过外推插值的办法估算出节假日的注入功率。具体方法如下。

在估算过程中分为两种情况：首先假设待计算的断面为第 n 年第 k 天的第 t 个断面，记 $P(n,k,t)$。

(1) 待计算的断面属于当天的前 10 个断面，即 $t \le 10$：则在估算时参照的是第 $n-1$ 年的前 m 天第 t 个断面的数据。

$$P(n,k,t) = \sum_{i=1}^{m} W_i \frac{P(n,k-i,t)}{P(n-1,k-i,t)} \times P(n-1,k,t) \quad (12)$$

式中：$P(n,k-i,t)$ 为第 n 年计算前 i 天的第 t 个断面；$P(n-1,k-i,t)$ 和 $P(n-1,k,t)$ 分别为第 n–1 年计算前 i 天与计算当天的第 t 个断面；W_i 为权重系数。

(2) 待计算的断面不属于当天的前 10 个断面，即 $t \ge 11$：则在估算时参照的是第 $n-1$ 年第 k 天的前 m 个断面数据。

$$P(n,k,t) = \sum_{i=1}^{m} W_i \frac{P(n,k,t-i)}{P(n-1,k,t-i)} \times P(n-1,k,t) \quad (13)$$

式中：$P(n,k,t-i)$ 为第 n 年第 k 天计算前的第 i 个断面；$P(n-1,k,t-i)$ 为第 n–1 年的第 k 天的第 t–i 与第 t 个断面；$P(n-1,k,t)$ 为第 n–1 年的第 k 天的第 t 个断面。

3.2.2 针对发生重大会议或者大型比赛的断面的修正

针对此类断面的修正思想主要是根据电量进行功率的估算。首先根据数据库中网架结构找到为发生特殊事件地区供电的具体线路，提取该线路所通过的电量，并由上一电压等级的系统图找到为该出线供电的节点。从而判断出了该特殊事件所影响的是哪个节点的负荷，即该节点的注入功率。由电量差可以计算功率差。计算公式为

$$P_{Mi} = [(P_{i-1} + P_i)/2 + \xi]/3.6 \times 10^6 \quad (14)$$

式中：P_{Mi} 为 i 时刻该节点的有功电度量；P_i, P_{i-1} 为 i, i–1 时刻该节点的有功功率；ξ 为准确系数。

4　基于关联规则修正的不良数据辨识

通过上述计算，可得到相对准确的节点注入功率预测修正值，将修正后的节点注入功率与待辨识断面的注入功率进行对比，偏差大于 2%(正常误差范围为 0.5%~2%)认为是不良数据，将其用预测修正值替换，偏差小于等于 2%的保留原来的测量值。最终可以得到待辨识断面相对准确的节点注入功率，然后针对所有的支路量测，将遥信量与遥测量进行相互校验，结果出现矛盾的量测量，经核实确定是遥测量有误后，应用与之最接近的断面数据来补充。

由于已经获得了相对准确的注入功率，以此为基础对其他量测进行辨识。将所有的支路量测值用式 $\left|Z_i-\sum_{j\in i}Z_{ij}\right|\le\varepsilon$ ($j\in i$ 指节点 j 与节点 i 相连)进行验证。将满足此约束的支路量测与相对准确的注入功率共同进行一次状态估计计算，估算出全部未参与状态估计计算的量测值，并求其残差，残差大于阈值的都属于不良数据，采用计算值进行替换，而小于阈值的则用其原断面的量测值。

以此方法辨识出的不良数据能够克服残差污染与残差淹没现象，因为在计算过程中采用的都是准确量测。该辨识方法的流程图如图 1 所示。

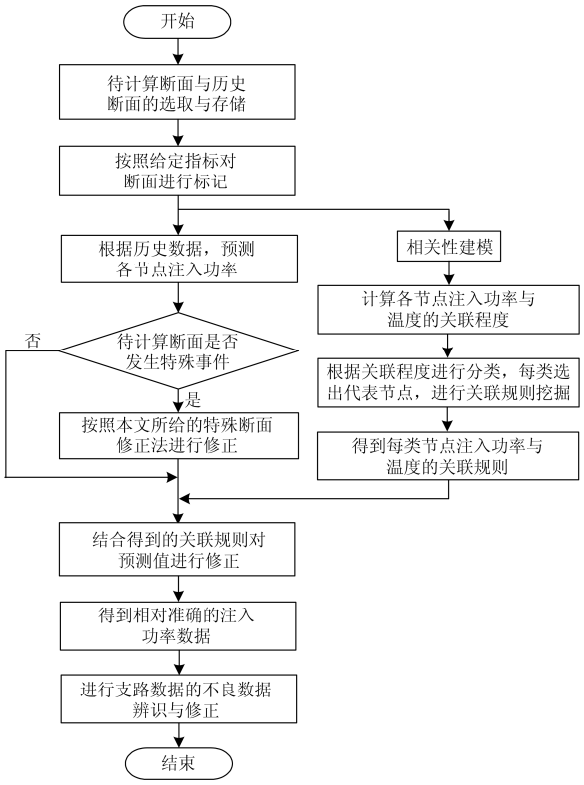

图 1　不良数据辨识流程图

Fig. 1 Flow chart of the bad data identification

5　算例仿真

本文以我国某一中等城市的实际输电网(主要是 220 kV 电压等级)为研究对象，采用本章所给方法进行不良数据的辨识。图 2 为该城市输电网系统的站外连线图，该连接图中：方框表示电源节点，双圆圈表示 500 kV 的厂站，而单圆圈表示 220 kV 厂站节点。

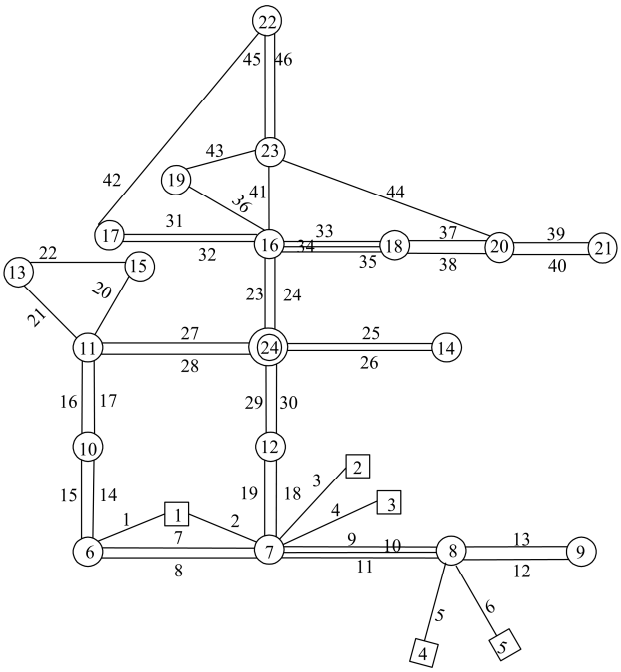

图 2　某中等城市系统接线图

Fig. 2 System diagram of a medium-sized city

5.1　关联度与关联规则挖掘仿真结果分析

首先根据已知的历史断面数据，计算出除平衡节点外的全部节点注入功率与气象因素的关联程度，由仿真结果可知，温度的关联程度最大，其他因素的影响可近似忽略。温度的关联度如表 1 所示。

表 1　各个节点注入功率与温度的相关度

Table 1 Correlation value between each node power injection and temperature

节点编号	相关度	节点编号	相关度	节点编号	相关度
1	0.5708	9	0.5125	17	0.4281
2	0.5649	10	0.5596	18	0.3199
3	0.4053	11	0.5779	19	0.3459
4	0.5314	12	0.4388	20	0.3354
5	0.5291	13	0.0923	21	0.3109
6	0.5583	14	0.1019	22	0.0698
7	0.4293	15	0.0892	23	0.3298
8	0.5011	16	0.3396	—	—

将 23 个节点的注入功率与温度间的关联关系分为五类进行讨论。如表 2 所示。

表 2 节点关联度的分类结果

Table 2 Classification of the nodes correlation values

类别(节点相关度区间)	节点编号
(0.55, 0.60]	1、2、6、10、11
(0.50, 0.55]	4、5、8、9
(0.40, 0.45]	3、7、12、17、
(0.30, 0.35]	16、18、19、20、21、23
(0.05, 0.15]	13、14、15、22

由表 2 可知，第五类，即节点 13、14、15、22 的注入功率变化幅度与温度关联不大，可以近似认为这 4 个节点的注入功率基本上不受温度的影响，挖掘过程中不考虑该类节点。根据本地区供电特点可知，第五类中的四个节点所供负荷有 90%以上为第一产业或第二产业，因此关联程度最小。而第一、二、三、四类的节点注入功率所带负荷为第一产业、第二产业、第三产业、农村用电或者城镇用电的混合负荷，只是每类产业所占比例不同。由本地区供电系统图可知，第一类节点的负荷中，第三产业及农村或城镇用电所占比例大于其他类别的节点负荷，因此从实际分析，也能够得出第一类节点的注入功率变化幅度与温度变化量的相关程度较大这一结论，从而说明了上述数据的合理性。

本文近似认为相关度相近的节点关联规则相近。因此，只在每类节点中各选一个节点作为代表进行关联规则的挖掘，所选节点分别为：节点 10、5、12、19。表 3 为每类节点关联规则的挖掘结果。

表 3 四类节点关联规则的挖掘结果

Table 3 Association rules mining results of the four types of nodes

节点编号	(01, 08]，(22, 24]	(08, 18]	(18, 22]
5	$\Delta T=1\Rightarrow\Delta P=[0.82\%,1.75\%]$	$\Delta T=1\Rightarrow\Delta P=[1.21\%,2.11\%]$	$\Delta T=1\Rightarrow\Delta P=[1.02\%,1.81\%]$
	$\Delta T=2\Rightarrow\Delta P=[1.75\%,3.01\%]$	$\Delta T=2\Rightarrow\Delta P=[2.59\%,4.12\%]$	$\Delta T=2\Rightarrow\Delta P=[2.65\%,3.53\%]$
	$\Delta T=3\Rightarrow\Delta P=[3.07\%,4.89\%]$	$\Delta T=3\Rightarrow\Delta P=[3.89\%,4.99\%]$	$\Delta T=3\Rightarrow\Delta P=[3.91\%,4.89\%]$
	$\Delta T=4\Rightarrow\Delta P=[4.25\%,5.67\%]$	$\Delta T=4\Rightarrow\Delta P=[5.87\%,6.58\%]$	$\Delta T=4\Rightarrow\Delta P=[5.56\%,6.75\%]$
10	$\Delta T=1\Rightarrow\Delta P=[0.75\%,1.72\%]$	$\Delta T=1\Rightarrow\Delta P=[1.62\%,2.85\%]$	$\Delta T=1\Rightarrow\Delta P=[1.28\%,2.37\%]$
	$\Delta T=2\Rightarrow\Delta P=[1.35\%,2.97\%]$	$\Delta T=2\Rightarrow\Delta P=[3.23\%,4.78\%]$	$\Delta T=2\Rightarrow\Delta P=[3.18\%,4.39\%]$
	$\Delta T=3\Rightarrow\Delta P=[2.58\%,4.17\%]$	$\Delta T=3\Rightarrow\Delta P=[5.67\%,6.70\%]$	$\Delta T=3\Rightarrow\Delta P=[4.48\%,6.25\%]$
	$\Delta T=4\Rightarrow\Delta P=[4.25\%,6.11\%]$	$\Delta T=4\Rightarrow\Delta P=[8.11\%,8.82\%]$	$\Delta T=4\Rightarrow\Delta P=[7.33\%,8.25\%]$
12	$\Delta T=1\Rightarrow\Delta P=[0.76\%,1.58\%]$	$\Delta T=1\Rightarrow\Delta P=[1.31\%,2.43\%]$	$\Delta T=1\Rightarrow\Delta P=[1.27\%,2.25\%]$
	$\Delta T=2\Rightarrow\Delta P=[1.48\%,2.41\%]$	$\Delta T=2\Rightarrow\Delta P=[3.01\%,4.63\%]$	$\Delta T=2\Rightarrow\Delta P=[2.87\%,4.19\%]$
	$\Delta T=3\Rightarrow\Delta P=[2.59\%,3.87\%]$	$\Delta T=3\Rightarrow\Delta P=[5.47\%,6.58\%]$	$\Delta T=3\Rightarrow\Delta P=[4.75\%,5.81\%]$
	$\Delta T=4\Rightarrow\Delta P=[3.57\%,5.02\%]$	$\Delta T=4\Rightarrow\Delta P=[7.80\%,8.84\%]$	$\Delta T=4\Rightarrow\Delta P=[6.55\%,7.79\%]$
19	$\Delta T=1\Rightarrow\Delta P=[0.77\%,1.88\%]$	$\Delta T=1\Rightarrow\Delta P=[1.28\%,2.69\%]$	$\Delta T=1\Rightarrow\Delta P=[1.23\%,2.01\%]$
	$\Delta T=2\Rightarrow\Delta P=[1.35\%,3.91\%]$	$\Delta T=2\Rightarrow\Delta P=[3.21\%,4.43\%]$	$\Delta T=2\Rightarrow\Delta P=[2.69\%,3.61\%]$
	$\Delta T=3\Rightarrow\Delta P=[2.91\%,3.89\%]$	$\Delta T=3\Rightarrow\Delta P=[4.97\%,6.25\%]$	$\Delta T=3\Rightarrow\Delta P=[4.42\%,5.25\%]$
	$\Delta T=4\Rightarrow\Delta P=[3.82\%,4.99\%]$	$\Delta T=4\Rightarrow\Delta P=[6.63\%,7.98\%]$	$\Delta T=4\Rightarrow\Delta P=[5.91\%,7.12\%]$

由表 3 中仿真结果可知，全天 24 h 的关联关系存在差异，可根据时间段不同将关联规则分为以上 3 类。以第一类为例进行说明，第一类为：(01, 08]，(22, 24]，即表明从凌晨 1 点到早上 8 点、从晚上 22 点到 24 点所挖掘得到的关联规则是相同的。而在此时间段内的关联规则又根据温差不同分为了 4 种情况即，$\Delta T=1$, $\Delta T=2$, $\Delta T=3$ 以及 $\Delta T=4$。不同的温差所引起的注入功率的变化幅度不同，此变化幅度的大小与该节点所带负荷的种类有密切关系。

结合表 2 与表 3 进行分析，在表 2 中可知第一类节点的关联度比较大。因此从理论上分析，节点 10 的关联规则中同样大小的温差所引起的节点注入功率变化的幅度要大于其他节点的变化幅度。再分析表 3，以第二列为例，显然节点 10 的变化幅度大于其他三类节点，从而说明该仿真结果与理论分析结果是一致的。置信度与支持度如表 4 所示。

支持度与置信度分别反映此关联规则的有效性和确定性，即支持度越高，关联程度越高；置信度越高，可信度越高。因为温差与功率变化幅度是弱相关，并且本文中每个节点的关联关系分为了四种

情况，因此，每种情况的支持度不可能太高，同时本文将所有温差的情况看作为一个整体，而上述四种情况只是其中部分情况，还存在$\Delta T=5$，$\Delta T=6$的情况，但其支持度都小于5%，所以在挖掘中没有涉及到此类小概率事件。由表 4 可知，每个时间段的四种关联规则的支持度的总和总是小于100%，并且接近100%的。结合本文实际情况，最小支持度阈值设定为 10%，最小置信度阈值为 60%，表 4 中小于阈值的关联规则认为不成立。

表 4 温差与功率变化幅度关联规则的置信度与支持度

Table 4 Confidence and support value of the association rules between the temperature difference and the power difference

节点编号	(01, 08]，(22, 24]		(08, 18)		(18, 22]	
	置信度/%	支持度/%	置信度/%	支持度/%	置信度/%	支持度/%
5	73.70	30.47	71.31	36.67	73.28	36.64
	81.94	24.13	75.88	23.39	68.86	20.04
	71.93	11.96	77.17	10.13	67.70	10.12
	65.17	11.86	78.05	11.51	57.69	11.90
10	84.08	34.29	79.16	40.48	89.55	44.78
	81.02	24.17	81.72	21.07	78.18	22.75
	69.50	16.20	74.47	14.11	76.99	13.81
	61.57	11.83	67.64	7.87	57.69	11.90
12	80.92	33.42	80.12	41.20	81.35	40.67
	72.80	30.03	74.24	19.07	74.32	21.63
	71.31	16.62	72.33	13.06	61.95	16.67
	75.51	14.32	59.35	12.70	50.00	15.48
19	73.07	29.80	72.61	43.93	79.89	39.95
	70.21	32.58	70.04	21.59	69.77	20.30
	60.37	14.07	62.92	13.99	61.06	16.43
	56.55	13.04	58.54	12.52	60.22	14.93

5.2 不良数据辨识的仿真结果分析

由于本文的重点是研究温差对注入有功功率变化幅度的关联规则，并结合此关联规则进行不良数据辨识，因此辨识的也主要是有功功率。以此为基础，为减少不必要的计算量，本章采用快速分解状态估计算法进行计算。

由 SCADA 系统所得的量测值如表 5 所示。由于该系统中存在大量双线路、三线路，而表 5 中的支路功率采用的是两个相连节点的编号，因此为了区分非单线路的量测值，采用如 P7-12,1 与 P7-12,2 的形式分别表示链接节点 7 与节点 12 的两条线路。

表 5 系统有功量测值

Table 5 Active power values

量测量	数值/MW	量测量	数值/MW	量测量	数值/MW	量测量	数值/MW
P1	220.92	P19	−446.77	P6-10,1	110.13	P16-17,2	172.57
P2	138.27	P20	−401.90	P6-10,2	88.03	P18-16,1	112.28
P3	137.02	P21	−122.16	P10-11,1	83.36	P18-16,2	95.18
P4	134.31	P22	−275.93	P10-11,2	75.10	P18-16,3	266.37
P5	129.77	P23	992.88	P7-12,1	2.22	P16-19	139.82
P6	−119.51	P1-6	111.69	P7-12,2	2.22	P19-16	−139.40
P7	−144.70	P1-7	109.23	P11-15	182.75	P18-20,1	187.40
P8	−151.50	P2-7	138.27	P11-13	291.95	P18-20,2	187.40
P9	−141.98	P3-7	137.02	P13-15	61.36	P20-21,1	61.08
P10	−39.70	P4-8	134.31	P15-13	−60.94	P20-21,2	61.08
P11	−141.40	P5-8	129.77	P24-16,2	155.39	P23-16	183.89
P12	−70.70	P7-6,1	102.99	P14-24,1	−58.53	P22-17	76.85
P13	−230.59	P6-7,1	−102.42	P24-14,1	58.77	P23-19	306.95
P14	−117.00	P8-7,1	−10.61	P24-11,1	230.94	P23-20	149.26
P15	−244.11	P8-7,2	−8.49	P24-11,2	226.70	P23-22,1	176.46
P16	−465.15	P8-7,3	−10.30	P12-24,1	−33.13	P22-23,1	−176.32
P17	−422.00	P8-9,1	72.30	P24-12,1	33.43	—	—
P18	848.64	P8-9,2	69.68	P16-17,1	172.57	—	—

根据文献[20]确定测量量的标准误差 δ，误差在 $\pm 3\delta$ 之外的测量可认为是不良数据，因此本文残差阈值取值为 2.81。待辨识断面为 2015 年 8 月 7 日上午 10 点时采集的数据。用于对待测断面进行修正的断面为：2015 年 8 月 7 日上午 9 点到 11 点的断面。

为便于仿真试验，假设表 5 中的量测量都是可靠数据，不存在不良数据。现将上述量测中 P8, P14, P14-24,1 与 P24-12,1 设置为不良数据，其值分别由 −151.50 MW 改为−158.00 MW，由−117.00 MW 改为 −105.00 MW， 由−58.53 MW 改为−50.79 MW， 由 33.43 MW 改为 40.97 MW。

经过快速解耦状态估计算法和标准残差辨识后，表 6 为标准残差 r_N 大于 2.81 的量测量。

表 6 大于 2.81 的量测量的标准残差 r_N

Table 6 r_N >2.81 of the measurements

量测	P8	P8-7,1	P8-7,2	P8-7,3	P14-24,1	P24-14,1
r_N	6.020	4.419	3.937	4.354	5.151	4.979

由表 6 得不良量测集为：P8，P8-7,1，P8-7,2，P8-7,3，P14-24,1，P24-14,1。因此可以断定，应用传统的标准残差法进行不良数据辨识时既发生了残差污染又发生了残差淹没现象，根据此表会将 P8-7,1，P8-7,2，P8-7,3，P24-14,1 误判定为不良数据，而将 P14 与 P24-12,1 误判为可靠量测。

接下来应用本文给出的方法进行辨识。辨识的过程主要分为以下几个部分：首先根据回归分析法对注入功率进行预测，然后应用关联规则与特殊事件处理方法对预测结果进行修正，进而根据修正后的预测结果对已知偏差大的注入功率数据进行修正。最后根据修正后的注入功率数据结合满足基尔霍夫定律的支路潮流进行状态估计计算与不良数据辨识。

应用回归分析法对各节点注入有功功率进行预测的预测结果如表 7 所示。

表 7 各节点注入功率预测值

Table 7 Injection power value of each node

节点编号	预测值/MW	节点编号	预测值/MW	节点编号	预测值/MW
1	217.84	9	−142.25	17	−422.73
2	143.67	10	−44.12	18	848.65
3	136.54	11	−142.49	19	−450.64
4	132.14	12	−70.36	20	−403.29
5	126.72	13	−230.00	21	−128.36
6	−115.04	14	−117.81	22	−276.25
7	−145.88	15	−244.70	23	993.61
8	−149.33	16	−464.40	—	—

以表 7 中所示的预测值为基础，应用关联规则对其进行修正。表 8 为初始预测值与经关联规则修正后的预测值取平均值后的修正结果。

图 3 为预测值与准确量测量差值和修正值与准确测量量差值的对比图。由此图可知经过关联规则修正后的预测值更加准确。

表 8 各节点注入功率预测值的修正结果

Table 8 Correction result value of each nodes' injection power

节点编号	预测值/MW	节点编号	预测值/MW	节点编号	预测值/MW
1	220.27	9	−140.60	17	−420.81
2	139.69	10	−38.57	18	849.93
3	137.36	11	−141.13	19	−447.04
4	134.98	12	−68.49	20	−401.96
5	129.86	13	−230.35	21	−122.64
6	−117.78	14	−117.43	22	−275.29
7	−143.98	15	−243.57	23	993.90
8	−149.60	16	−464.78	—	—

图 3 预测值修正前后的误差对比图

Fig. 3 Error comparison chart of the predictive value before and after the correction

将修正后的节点注入功率与已知断面的注入功率(添加不良数据后的断面)进行对比，偏差大于 2% 的认为是不良数据，将其用预测修正值替换，偏差小于等于2%的保留原来的测量值。最终被替换的注入功率有 P8 与 P14，分别由原来的−158.00 MW 被替换为−149.60 MW，由原来的−105.00 MW 被替换为−117.43 MW。以此修正结果为基础，应用基尔霍夫定律判断支路潮流中较为准确的量测。以 P1、P6 为例来说明判断过程。

P1=220.92 MW.P(1-6)+P(1-7)=111.69+109.25=220.94 MW，220.92 与 220.94 的差值小于 P1 数值的 2%，因此满足基尔霍夫定律，初步判断 P(1-6)与 P(1-7) 为可靠量测；应用基尔霍夫定律可知 P6=P(6-10,1)+P(6-10,2)+P(6-1)+P(6-7,1)+P(6-7,2)，但是 P(6-1)与 P(6-7,2)都不是已知量测，因此与节点 6 相连接的支路都不能够证明其可靠性。依次类推，最终得出部分可靠量测，并应用快速解耦法进行状态估计计算，计算各个量测的残差，此步骤辨识的是支路潮流的量测，而上述负荷预测与关联规则结

合所辨识的是注入功率的量测。最终得到大于 2.81 的支路量测量的标准残差 r_N，如表 9 所示。

表 9　r_N 大于 2.81 的支路量测

Table 9　r_N >2.81 of the branch measurements

量测量	P24-12,1	P14-24,1
r_N	4.820	5.430

结合表 6 与表 9 进行分析可知，本文给出的辨识不良数据的方法能够克服残差污染以及残差淹没现象。准确地辨识出了 P8, P14, P14-24,1 与 P24-12, 1 共 4 个不良数据，从而证明了本文所提方法的有效性。

6　结论

本文提出一种基于负荷预测与关联规则修正的不良数据辨识方法。该方法首先根据数据量之间的内在联系，给出了一种三维矩阵的数据存储方法，建立基于回归分析法的预测模型与基于灰色关联的相关性分析模型，并采用关联规则与特殊断面修正法对预测值进行修正，进而完成对注入功率的辨识。在此基础上，再通过基尔霍夫定律与残差辨识法完成对支路潮流数据的辨识工作。仿真算例证明了该方法能够在克服残差污染和残差淹没现象的前提下准确辨识出全部不良数据。

参考文献

[1]　于尔铿. 电力系统状态估计[M]. 北京: 水利电力出版社, 1985.

[2]　黄知超, 谢霞, 王斌. 结合模糊综合评判与决策的电力系统状态估计[J]. 电力系统保护与控制, 2015, 43(7): 65-69.
HUANG Zhichao, XIE Xia, WANG Bin. Power system state estimation combined with fuzzy comprehensive evaluation and decision-making[J]. Power System Protection and Control, 2015, 43(7): 65-69.

[3]　丁宏恩, 戴则梅, 霍雪松, 等. 基于混合量测的二次线性状态估计方法及其工程应用[J]. 电力系统保护与控制, 2016, 44(13): 163-170.
DING Hongen, DAI Zemei, HUO Xuesong, et al. Hybrid measurement based secondary linear state estimation method and its engineering application[J]. Power System Protection and Control, 2016, 44(13): 163-170.

[4]　孟建良, 刘德超. 一种基于 Spark 和聚类分析的辨识电力系统不良数据新方法[J]. 电力系统保护与控制, 2016, 44(3): 85-91.
MENG Jianliang, LIU Dechao. A new method for identifying bad data of power system based on Spark and clustering analysis[J]. Power System Protection and Control, 2016, 44(3): 85-91.

[5]　赵海天, 相年德, 王世缨, 等. 多不良数据的相关量测检测方法[J]. 中国电机工程学报, 1990, 10(6): 24-30.
ZHAO Haitian, XIANG Niande, WANG Shiying, et al. A new detection method for bad date using linear dependence properties of the measurements in power system state estimation[J]. Proceedings of the CSEE, 1990, 10(6): 24-30.

[6]　刘浩, 崔巍. *R*n 检测与状态预估相结合的不良数据检测辨识法[J]. 电力系统及其自动化学报, 2001, 13(2): 39-42.
LIU Hao, CUI Wei. The detection and identification method of bad data combined *R*n detection and state forecast[J]. Proceedings of the CSU-EPSA, 2001, 13(2): 39-42.

[7]　倪小平, 张步涵. 基于等效电流量测变换的状态估计及不良数据检测与辨识方法[J]. 电网技术, 2002, 26(8): 12-15.
NI Xiaoping, ZHANG Buhan. A state estimation method for dad data detection and identification based on equivalent current measurement transformation[J]. Power System Technology, 2002, 26(8): 12-15.

[8]　郭艳东, 申定辉. 基于改进 GSA 算法的电网不良数据检测与辨识[J]. 华东电力, 2013, 41(3): 542-545.
GUO Yandong, SHEN Dinghui. Bad data detection and identification based on improved GSA algorithm in grid[J]. East China Electric Power, 2013, 41(3): 542-545.

[9]　孙国强, 卫志农, 周封伟. 改进迭代自组织数据分析法的不良数据辨识[J]. 中国电机工程学报, 2006, 26(11): 162-166.
SUN Guoqiang, WEI Zhinong, ZHOU Fengwei. The application of ISODATA to bad data detection and identification based on genetic algorithms[J]. Proceedings of the CSEE, 2006, 26(11): 162-166.

[10]　卢志刚, 张宗伟. 基于量测量替换与标准化残差检测的不良数据辨识[J]. 电力系统自动化, 2007, 31(13): 52-56.
LU Zhigang, ZHANG Zongwei. Bad data identification based on measurement replace and standard residual detection[J]. Automation of Electric Power Systems, 2007, 31(13): 52-56.

[11]　陈强, 金小明, 姚建刚, 等. 基于改进 ABC 算法的中长期电力负荷组合预测[J]. 电力系统保护与控制, 2014, 42(23): 113-117.
CHEN Qiang, JIN Xiaoming, YAO Jiangang, et al. Improved artificial bee colony algorithm applied to medium and long-term load combination forecasting[J].

Power System Protection and Control, 2014, 42(23): 113-117.

[12] FAY D, RINGWOOD J V. On the influence of weather forecast errors in short-term load forecasting models[J]. IEEE Transactions on Power Systems, 2010, 25(3): 1751-1758.

[13] 刘念, 张清鑫, 刘海涛. 基于核函数极限学习机的微电网短期负荷预测方法[J]. 电工技术学报, 2015, 30(8): 218-224.
LIU Nian, ZHANG Qingxin, LIU Haitao. Online short-term load forecasting based on ELM with kernel algorithm in micro-grid environment[J]. Transactions of China Electrotechnical Society, 2015, 30(8): 218-224.

[14] 高赐威, 李倩玉, 苏卫华, 等. 短期负荷预测中考虑积温效应的温度修正模型研究[J]. 电工技术学报, 2015, 30(4): 242-248.
GAO Ciwei, LI Qianyu, SU Weihua, et al. Temperature correction model research considering temperature cumulative effect in short-term load forecasting[J]. Transactions of China Electrotechnical Society, 2015, 30(4): 242-248.

[15] 李龙, 魏靖, 黎灿兵, 等. 基于人工神经网络的负荷模型预测[J]. 电工技术学报, 2015, 30(8): 225-230.
LI Long, WEI Jing, LI Canbing, et al. Prediction of load model based on artificial neural network[J]. Transactions of China Electrotechnical Society, 2015, 30(8): 225-230.

[16] 姜玉山, 刘嘉婧, 卢毅. 基于回归分析的频域分解短期负荷预测方法[J]. 电力需求侧管理, 2011, 13(5): 12-15.
JIANG Yushan, LIU Jiajing, LU Yi. Short-term load forecasting based on frequency analysis of regression[J]. Power Demand Side Management, 2011, 13(5): 12-15.

[17] 柏丹丹, 和敬涵, 王小君, 等. 基于灰色关联分析的自适应低频减载[J]. 电力系统保护与控制, 2014, 42(6): 20-25.
BAI Dandan, HE Jinghan, WANG Xiaojun, et al. Adaptive UFLS scheme based on grey correlation analysis[J]. Power System Protection and Control, 2014, 42(6): 20-25.

[18] 肖峻, 张晶, 朱涛, 等. 基于关联分析的城市用电负荷研究[J]. 电力系统自动化, 2007, 31(17): 103-107.
XIAO Jun, ZHANG Jing, ZHU Tao, et al. Analysis of urban power load based on association rules[J]. Automation of Electric Power Systems, 2007, 31(17): 103-107.

[19] 谢龙君, 李黎, 程勇, 等. 融合集对分析和关联规则的变压器故障诊断方法[J]. 中国电机工程学报, 2015, 35(2): 277-286.
XIE Longjun, LI Li, CHENG Yong, et al. A fault diagnosis method of power transformers by integrated set pair analysis and association rules[J]. Proceedings of the CSEE, 2015, 35(2): 277-286.

[20] 李碧君, 薛禹胜, 顾锦汶, 等. 基于权函数的电力系统状态估计算法[J]. 电力系统自动化, 1999, 23(8): 32-34.
LI Bijun, XUE Yusheng, GU Jinwen, et al. Weight function based power system state estimation algorithm[J]. Automation of Electric Power Systems, 1999, 23(8): 32-34.

收稿日期：2016-11-06；　　修回日期：2017-01-21

作者简介：

王海云(1988—)，女，硕士，工程师，研究方向为电网仿真、线损计算、分布式电源接入；E-mail: 13520396759@139.com

田莎莎(1989—)，女，硕士，研究方向为电力系统状态估计。E-mail: tianshasha0317@163.com

(编辑　周金梅)