Consumer-branch Connectivity Identification of Low Voltage Distribution Networks Based on Data-driven Approach

Yongjun Zhang, Yingqi Yi, Wenyang Deng, Siliang Liu, Lai Zhou, Kaidong Lin, and Yongzhi Cai

***Abstract*—Accurate topological information is crucial in supporting the coordinated operational requirements of source-load-storage in low-voltage distribution networks. Comprehensive coverage of smart meters provides a database for low-voltage topology identification (LVTI). However, because of electricity theft, power line communication crosstalk, and interruption of communication, the measurement data may be distorted. This can seriously affect the performance of LVTI methods. Thus, this paper defines hidden errors and proposes an LVTI method based on layer-by-layer stepwise regression. In the first step, a multi-linear regression model is developed for consumer-branch connectivity identification based on the energy conservation principle. In the second step, a significance factor based on the t-test is proposed to modify the identification results by considering the hidden errors. In the third step, the regression model and significance threshold parameters are iteratively updated layer by layer to improve the recall rate of the final identification results. Finally, simulations of a test system with 63 users are carried out, and the practical application results show that the proposed method can guarantee over 90% precision under the influence of hidden errors.**

***Index Terms*—Data driven, hidden error, linear regression, low voltage distribution network, topology identification.**

Received: September 11, 2023
Accepted: January 21, 2024
Published Online: July 1, 2024

Yongjun Zhang, Wenyang Deng, Siliang Liu, Kaidong Lin are with the School of Electric Power Engineering and Guangdong Key Laboratory of Clean Energy Technology, South China University of Technology, Guangzhou 510641, China (e-mail: zhangjun@scut.edu.cn; dwyang@scut.edu.cn; liang_in_ps@163.com; 596875439@qq.com).

Yingqi Yi (corresponding author) is with the School of Electric Power Engineering, South China University of Technology, Guangzhou 510641, China (yi_yingqi@foxmail.com).

Lai Zhou is with the Guangzhou Panyu vocational and technical college, Guangzhou 511483, China (e-mail: 291666680@qq.com).

Yongzhi Cai is with the Metrology Center Guangdong Power Company Grid, Qingyuan 511500, China (e-mail: dkycyz@163.com).

DOI: 10.23919/PCMP.2023.000465

NOMENCLATURE

A. Abbreviations

LVDN	low voltage distribution network
LVTI	low-voltage topology identification
CBCI	consumers-branch connectivity identification
PLC	power line communication
LSR	layer-by-layer stepwise regression
SR	stepwise regression
RN	root node
SN	stem node
LN	leaf node
ETEs	electricity theft errors
PCEs	PLC crosstalk errors
CMEs	communication mistake errors
LS	least squares
IQP	integer quadratic programming
LASSO	least absolute shrinkage and selection operator

B. Variables

J	set of indices of stem nodes
H	set of indices of measurements
C	set of indices of leaf nodes
Q	set of electric larceny consumers
Z	set of non-segment consumers
K	set of consumers with communication anomalies
X_{Φ}	subset of leaf nodes depended on the Φth stem node
ξ_{Φ}	subset of significant factor corresponding to X_{Φ}
$X_{\#}$	intersection of subset of leaf nodes depended on different stem node
$\xi_{\#\max}$	set of significant factor maximum corresponding to $X_{\#}$
$X_{\mathrm{R}\Phi}$	corrected subset of leaf nodes depended on the Φth stem node
C_{Φ}	set of indices of leaf nodes in subset X_{Φ}

$C_{\#}$	set of indices of leaf nodes in intersection $X_{\#}$
$C_{\mathrm{R}\Phi}$	set of indices of leaf nodes in corrected subset
$\tilde{\boldsymbol{I}}_{\Phi}$	vector of injection current phasor for the Φth stem node
$\boldsymbol{I}_{\Phi}$	vector of injection current magnitude for the Φth stem node
$\boldsymbol{\omega}_{\Phi}$	vector of coefficient for the Φth stem node
$\tilde{\boldsymbol{I}}_{\mathrm{D}}$	matrix of current phasor for leaf nodes
$\boldsymbol{I}_{\mathrm{D}j}$	vector of injection current magnitude for the jth leaf node
$\boldsymbol{Y}$	vector of the current magnitude measurements for the stem node
$\boldsymbol{X}$	design matrix of the current magnitude measurements for the leaf nodes
$\boldsymbol{\beta}$	vector of regression coefficient
$\boldsymbol{e}$	vector of errors
$\boldsymbol{\beta}^{*}$	least squares estimation of regression coefficients
$\boldsymbol{e}_{\mathrm{s}}$	vector of measurement errors
$\boldsymbol{e}_{\mathrm{h}}$	vector of hidden errors
$\boldsymbol{e}_{\mathrm{hq}}$	vector of electricity theft errors
$\boldsymbol{e}_{\mathrm{hz}}$	vector of PLC subnetwork information errors
$\boldsymbol{e}_{\mathrm{hk}}$	vector of communication mistake errors
$\tilde{\boldsymbol{I}}_{\Phi i}$	injection current phasor of the Φth stem node at the ith time instant
$\tilde{\boldsymbol{I}}_{\mathrm{D}ij}$	load current phasor of the jth leaf node at the ith time instant
$I_{\mathrm{D}ij}$	load current magnitude of the jth leaf node at the ith time instant
$I_{\mathrm{DQ}ij}$	measurement of load current for the jth electric larceny consumer at the ith time instant
β_j	regression coefficient of the jth independent variable
P_0	P-value for the t-test of $\beta_j = 0$
P_1	P-value for the t-test of $\beta_j = 1$
ξ	significance factor
$x_{\Phi}(j)$	leaf node corresponding to the jth significant independent variable in subset X_{Φ}
$\xi_{\Phi}(j)$	significance factor for $x_{\Phi}(j)$
$x_{\#}(j)$	leaf node corresponding to the jth significant independent variable in intersection X_{Φ}
$\xi_{\#}(j)$	maximum significance factor for $x_{\#}(j)$ in the different X_{Φ}
$x_{\mathrm{R}\Phi}(j)$	leaf node corresponding to the jth significant independent variable in correct subset $X_{R\Phi}$
n_{Φ}	number of the significant independent variables in subset X_{Φ}
$n_{\#}$	number of the significant independent variables in subset $X_{\#}$
$n_{\mathrm{R}\Phi}$	number of the significant independent variables in correct subset $X_{\mathrm{R}\Phi}$
$\boldsymbol{\Omega}_{\mathrm{p}}$	precision rate
$\boldsymbol{\Omega}_{\mathrm{r}}$	recall rate
N_{correct}	number of consumers with identifiable phase connectivity information from algorithms
N_{output}	number of consumers with correct phase identification from the outputs of algorithms
λ_{entry}	significance introduced threshold
$\lambda_{\mathrm{remove}}$	significance remove threshold
$\Delta\lambda$	increment of significance threshold in each iteration
$\lambda_{\max}$	maximum significance threshold.

I. Introduction

As the climate goals of "carbon emission peak" and "carbon neutrality" proposed by China are increasingly being valued by various countries [1] and a large number of end-users are actively accelerating power substitution and emission reduction transformation, the low-voltage distribution network (LVDN) is expected to meet the requirements of flexible configuration and coordinated operation of source-load-storage resource in the future [2]. The realization of these goals is highly dependent on the digitalization level of the LVDN, and one of its important features is complete and accurate topology information. However, because of the limited coverage of measurement devices, complex network structure, and frequent changes, the network topology information of an LVDN has generally been unavailable or inaccurate for a long time. This has limited the intelligence and refinement of LVDN planning, operation, and management [3].

Research on network topology identification has been carried out from a very early stage. Initially, topology identification was carried out to solve the problem of possible false alarms in remote telecommunications [4]. Subsequently, the concept appeared in power system state estimation studies, though the identification was mainly for transmission grids [5]. In recent years, with the comprehensive coverage of the advanced measurement system in LVDN, the low-voltage topology identification (LVTI) problem has gradually attracted widespread attention. The existing LVTI methods that are based on being data-driven can be divided into two categories: one based on voltage correlation and the other on energy conservation. These methods involve correlation analysis, dimension reduction, linear regression, integer programming, maximum likelihood estimation, machine learning, etc.

Based on voltage correlation, references [6] and [7] correct the LVDN topology connection information by comparing correlation of the time series of voltage measurements. However, the performance can be poor in the event of short electrical distances, light load, or balanced three-phase load. References [8] and [9] use the linear regression method to identify node connectivity and line parameters based on the line voltage drop model, while reference [10] extends its application to three-phase four-wire LVDN. However, they are restricted to only finding parent nodes because of the radial topology assumption. Reference [11] proves that the node connectivity identification problems, in radial and mesh structures, can be formulated as a linear regression with group lasso. In [12] and [13], a nodal voltage probability distribution model and a Markov random field model are established, respectively, to realize topology identification with maximum-likelihood estimation. For machine learning, reference [14] compares the performance of supervised learning methods such as decision trees, random forests, and AdaBoost algorithms. As the training data labels are difficult to obtain in practice, the application of such methods is limited. For unsupervised learning, the methods for LVTI based on a clustering algorithm are proposed in [15]. Although it does not need training data, it is sensitive to the algorithm parameters and this may lead to serious errors.

Based on energy conservation, an integer quadratic programming model is established in [16], [17] for phase identification. However, it is sensitive to bad and incomplete data. Reference [18] converts the LVTI problem into a mixed-integer linear program. Although the efficiency of the solution is improved, it requires the acquisition of phase angle information, which in turn requires the installation of phasor measurement units and is a very costly investment for LVDN [11]. In [19], LVTI is transformed into a maximum-likelihood estimation model of system parameter values, and an iterative solution method is used to estimate the line parameters and topology structure. Based on the study, the influence of measurement error and network topology change is taken into consideration, and a more general method is proposed in [20]. However, it is necessary to have measurement devices at each node, which is not applicable in LVDN. References [21] and [22] compress the dimension of current measurements of the upstream and downstream nodes for topology identification based on the Fourier transform or principal component analysis [23]. Reference [24] proposes a phase identification method based on lasso regression, though it only considers the effect of measurement noise on the identification performance. Considering the hidden error, reference [25] proposes a stepwise regression (SR) algorithm to improve the precision of consumer phase identification. However, high precision accuracy is obtained at the expense of recall, and the exact definition of hidden error is lacking.

Building on the research of [25], an layer-by-layer stepwise regression (LSR) algorithm is proposed in this paper, based on the SR model and the iterative updating of significance threshold parameters. The proposed method uses only the current measurements to quantitatively analyze the influence of different types of hidden error, such as electricity theft [26], power line communication (PLC) crosstalk [27], and interruption of communication [28]. The main contributions of this paper are:

1) Based on the principle of energy conservation, an LSR algorithm for consumer-branch connectivity identification (CBCI) is proposed to improve the recall rate compared with [25].

2) The three most common types of hidden error are defined and methods for quantitative assessment of hidden errors in LVDN are proposed.

3) The identification performance of the LSR algorithm with different types and sizes of hidden error scenarios is analyzed by comparison with other methods.

4) The LSR algorithm is applied to identify actual LVDN topology, and the feasibility of the proposed method is verified by comparison with real topology data after field investigation.

The remainder of this paper is organized as follows. Section II presents the problem discription of consumer-branch connectivity identification in LVDN. Section III defines hidden errors, and Section IV proposes an LVTI method based on layer-by-layer stepwise regression. In Section V, simulation results of proposed method is presented, and Section VI presents the main conclusions of the paper.

II. Problem Description

LVDN is located at the end of the power system and can be structurally understood as a collection of several power supply units with an middle voltage/ low voltage (MV/LV) transformer as the root node. This can be called the LVDN segment. An LVDN segment is a relatively independent power supply network consisting of an MV/LV transformer, power lines, and consumers. As shown in Fig. 1, the network topology of an LVDN segment can be described as a radial tree structure consisting of a root node (RN), stem nodes (SNs) and leaf nodes (LNs). RN represents the MV/LV transformer, SNs represent the A/B/C phase feeds on the LV side of the MV/LV transformer or the branches, and LNs represent the equivalent single-phase consumers, where a three-phase consumer can be equal to three single-phase consumers. From the power flow perspective, when the integration of distributed generations is not considered, the energy consumption of all LNs in the LVDN is supplied by a specific SN upstream.

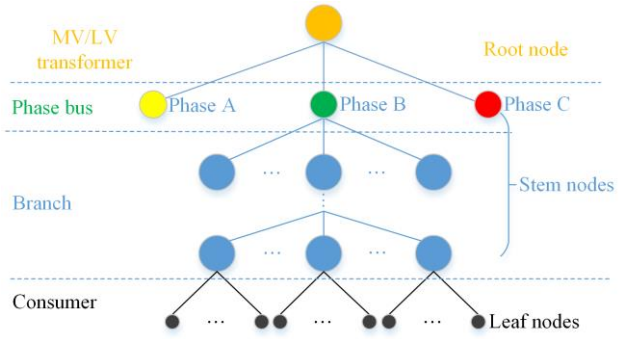


Fig. 1. LVDN topology.

The LVTI described in this paper is to identify the consumer-branch connectivity in the LVDN segment. This can also be called the stem-leaf node connectivity relationship. In general, the root-leaf node connectivity relationship can be obtained based on the PLC sub-network information stored in the data concentrator units [27]. As shown in Fig. 1, all LNs are connected to the SN that represents the B phase feed.

At present, smart meters or current sensors are installed in almost all LNs and partial SNs to collect current and voltage measurement data [30] with a resolution of 15 min. Suppose there are N LNs and M SNs in this LVDN segment, as well as T measurements of current data. Let $J=\{1,2,\cdots,M\}$, $H=\{1,2,\cdots,T\}$, and $C=\{1,2,\cdots,N\}$. The SNs current vector is $\tilde{\boldsymbol{I}}_{\Phi}\in\mathbb{R}^{T}$, the LNs load current matrix is $\tilde{\boldsymbol{I}}_{\mathrm{D}}\in\mathbb{R}^{T\times N}$. The topology coefficient vector $\boldsymbol{\omega}_{\Phi}\in\mathbb{R}^{N}$ can be expressed as:

$$\tilde{\boldsymbol{I}}_{\Phi}=[\tilde{\boldsymbol{I}}_{\Phi 1},\cdots\tilde{\boldsymbol{I}}_{\Phi i},\cdots,\tilde{\boldsymbol{I}}_{\Phi T}],\ i\in H,\ \Phi\in J \tag{1}$$

$$\tilde{\boldsymbol{I}}_{\mathrm{D}}=[\tilde{\boldsymbol{I}}_{\mathrm{D}ij}]_{(T\times N)},\ i\in H,\ j\in C \tag{2}$$

$$\boldsymbol{\omega}_{\Phi}=[\boldsymbol{\omega}_{\Phi 1},\cdots,\boldsymbol{\omega}_{\Phi j}\cdots,\boldsymbol{\omega}_{\Phi N}]^{\mathrm{T}},\ j\in C,\ \Phi\in J \tag{3}$$

where J is the set of indices of SNs; H is the set of indices of measurements; C is the set of indices of LNs; $\boldsymbol{\omega}_{\Phi j}=\{0,1\}$ denotes the connectivity relationship of the jth LN to the Φth SN, with 1 indicating they are connected and 0 that they are not; $\tilde{\boldsymbol{I}}_{\Phi i}$ is the injection current phasor of the Φth SN at the ith time instant; and $\tilde{\boldsymbol{I}}_{\mathrm{D}ij}$ is the load current phasor of the jth LN at the ith time instant. From Kirchhoff's current law, these linear equations can be expressed as:

$$\tilde{\boldsymbol{I}}_{\Phi}=\tilde{\boldsymbol{I}}_{\mathrm{D}}\boldsymbol{\omega}_{\Phi} \tag{4}$$

LVTI is primarily based on solving (4) to obtain the topology coefficient vector, which reflects the relationship of connectivity between SNs and LNs. Since the phase angle data cannot be measured by sensors or smart meters [29], current magnitude measurements are used to replace the current phasor measurements. Considering the impact of errors, the current magnitude measurements can be expressed as:

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e} \tag{5}$$

$$\boldsymbol{e}=\boldsymbol{e}_{\mathrm{s}}+\boldsymbol{e}_{\mathrm{h}} \tag{6}$$

where $\boldsymbol{Y}\in\mathbb{R}^{T}$ is the vector of the current magnitude measurements for the SN; $\boldsymbol{X}\in\mathbb{R}^{T\times N}$ is the design matrix of the current magnitude measurements for the LNs; $\boldsymbol{\beta}=[\beta_1,\cdots,\beta_j\cdots,\beta_N]^{\mathrm{T}}$ is the regression coefficient vector with unknown values; the errors $\boldsymbol{e}\in\mathbb{R}^{T}$ is a random error vector, including the measurement errors $\boldsymbol{e}_{\mathrm{s}}\in\mathbb{R}^{T}$ and hidden errors $\boldsymbol{e}_{\mathrm{h}}\in\mathbb{R}^{T}$. The measurement errors $\boldsymbol{e}_{\mathrm{s}}$ are caused by meter reading errors and clock synchronization errors, and so can be modelled to be a zero-mean Gaussian distribution [23]. The measurement error rate ε_{s} ranges from 0.1% to 8% considering metering errors and clock synchronization errors simultaneously [25].

III. Hidden Errors

Measurement errors can be predicted in size, and are referred to in this paper as "apparent errors". In addition, considering that there may be serious distortion of measurement data due to severe problems such as electricity theft, PLC crosstalk and communication interruptions in LVDN, the errors caused by such problems are generally unpredictable and hidden, so they are called "hidden errors" in this paper. The three types of hidden errors are: electricity theft errors (ETEs) $\boldsymbol{e}_{\mathrm{hq}}$, PLC crosstalk errors (PCEs) $\boldsymbol{e}_{\mathrm{hz}}$ and communication mistake errors (CMEs) $\boldsymbol{e}_{\mathrm{hk}}$. The total hidden errors $\boldsymbol{e}_{\mathrm{h}}$ is therefore given as:

$$\boldsymbol{e}_{\mathrm{h}}=\boldsymbol{e}_{\mathrm{hq}}+\boldsymbol{e}_{\mathrm{hz}}+\boldsymbol{e}_{\mathrm{hk}} \tag{7}$$

ETEs refer to hidden errors introduced by electricity theft, which is usually caused by modifying smart meter records or bypassing electricity consumption in such a way that the measurements of the customer's load current are zero or much smaller than the real load current. In a certain time period S, the magnitude and rate of ETEs can be expressed respectively as:

$$\boldsymbol{e}_{\mathrm{hq}}=\frac{1}{S}\sum_{i\in S}\left[\sum_{j\in Q}(I_{\mathrm{D}ij}-I_{\mathrm{DQ}ij})\right] \tag{8}$$

$$\varepsilon_{\mathrm{hq}}=\frac{1}{S}\sum_{i\in S}\left[\sum_{j\in Q}(I_{\mathrm{D}ij}-I_{\mathrm{DQ}ij})/\sum_{j\in N}I_{\mathrm{D}ij}\right] \tag{9}$$

where Q is the set of electric larceny consumers; $I_{\mathrm{DQ}ij}$ is the measurement of load current for the jth electric larceny consumer at the ith time instant; and $I_{\mathrm{D}ij}$ is the true value of load current magnitude for the jth LN at the ith time instant. Considering that the phase angle data cannot be measured by sensors or smart meters in an LVDN, the current magnitude $I_{\mathrm{D}ij}$ is used in practice to replace the current phasor $\tilde{\boldsymbol{I}}_{\mathrm{D}ij}$.

PCEs refer to hidden errors introduced by PLC crosstalk between neighboring LVDN segments, resulting in PLC subnetwork information that reflects the incorrect root-leaf node connection relationship of this LVDN segment. This relationship is manifested by the mixing of non-subscribers in the user file of this segment. The load current data collected by non-subscribers of this segment seriously interfere with the identification of the stem-leaf node dependency relationship of this LVDN segment. In a certain period *S*, the magnitude and rate of PCEs can be expressed respectively as:

$$\boldsymbol{e}_{\mathrm{hz}}=\frac{1}{S}\sum_{i\in S}\left(\sum_{j\in \mathrm{Z}} I_{\mathrm{D}ij}\right) \tag{10}$$

$$\varepsilon_{\mathrm{hz}}=\frac{1}{S}\sum_{i\in S}\left(\sum_{j\in \mathrm{Z}} I_{\mathrm{D}ij}/\sum_{j\in N} I_{\mathrm{D}ij}\right) \tag{11}$$

where *Z* is the set of non-segment consumers.

CMEs refer to hidden errors introduced by external interference or relay anomalies during the consumer load data collection process, resulting in missing or zero load data for some consumers at some time instants. The magnitude and rate of CMEs within a certain time period *S* can be expressed respectively as:

$$\boldsymbol{e}_{\mathrm{hk}}=\frac{1}{S}\sum_{i\in S}\left(\sum_{j\in K} I_{\mathrm{D}ij}\right) \tag{12}$$

$$\varepsilon_{\mathrm{hk}}=\frac{1}{S}\sum_{i\in S}\left(\sum_{j\in K} I_{\mathrm{D}ij}/\sum_{j\in N} I_{\mathrm{D}ij}\right) \tag{13}$$

where *K* is the set of consumers with communication anomalies.

The total hidden error rate ε_{h}, when ETEs, PCEs, and CMEs are present simultaneously in the segment area, can be calculated as:

$$\varepsilon_{\mathrm{h}}=\varepsilon_{\mathrm{hq}}+\varepsilon_{\mathrm{hz}}+\varepsilon_{\mathrm{hk}} \tag{14}$$

IV. Topology Identification Method

The SR method proposed in [25] provides a systematic way of identifying the network topology in a statistical framework that takes hidden errors into account. However, as the results of the SR method are easily affected by the parameter settings of the algorithm, it can lead to partial stem-leaf node connectivity relationships which cannot be determined. Thus, an LSR method is proposed in this paper and its key steps are as follows.

1) A multiple linear regression model shown in (5) is formulated using the current magnitudes of SN and LNs as a dependent variable and independent variables, respectively.

2) The significance threshold parameters are set, and the SR algorithm is used to obtain each stem-leaf node subset that reflects the connection relationship between SNs and LNs, while the results are corrected based on the significance factor.

3) The independent variables for which the subset of stem-leaf node has been specified in the previous step are removed from the set of LNs, and the SR calculation is continued after updating the regression model and increasing the significance threshold values.

4) Step 3 is repeated until the significance threshold reaches the maximum limit and the final identification result is specified.

A. Significance of Regression Coefficients

1) Estimation of Regression Coefficients

For the linear regression model (5), if errors $\boldsymbol{e}$ satisfies the normal distribution, i.e. $\boldsymbol{e}\sim N(0,\sigma^2\boldsymbol{I})$, the least squares estimation of regression is given as:

$$\boldsymbol{\beta}^*\sim N(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \tag{15}$$

$$\beta_j^*\sim N(\beta_j,\sigma^2 c_{jj}) \tag{16}$$

where $\boldsymbol{\beta}^*$ is the least squares estimation of regression coefficients; σ is the standard deviation of the residuals of a fitted linear regression model; c_{jj} is the diagonal elements of matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Equation (15) indicates that $\boldsymbol{\beta}^*$ is an unbiased estimation of regression coefficients, which can be interpreted as that this estimation has no systematic bias. It means that the estimation of regression coefficients may sometimes be lager or small in different sample spaces, but these deviations, which can be positive or negative, are statistically equal to zero on average.

According to (16), the error of the estimated regression coefficient β_j^* can be expressed as:

$$\mathrm{Var}(\beta_j^*)^{1/2}=\sigma^*\sqrt{c_{jj}} \tag{17}$$

where $\mathrm{Var}(\beta_j^*)^{1/2}$ reflects the variation range of different sample spaces. As the value of σ is generally unknown, an unbiased estimate σ^* of σ is used instead, i.e.:

$$\sigma^*=\sqrt{S_{\mathrm{SE}}/(S-N)} \tag{18}$$

$$S_{\mathrm{SE}}=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{19}$$

where *SSE* is the residual sum of squares, whose size reflects the degree of fitting between the actual data and the theoretical model in (5). The smaller the *SSE*, the better the fitting between the data and the model.

If $\mathrm{Var}(\beta_j^*)^{1/2}$ is small, the least squares estimation of the regression coefficient can be considered to be more precise. Thus, the relationship of connectivity between SNs and LNs can be determined based on the estimation of the regression coefficient. However, errors, and especially the hidden errors, can lead to a large *SSE*, and thus, the standard deviation of the estimation of regression coefficients can be large. It indicates that estimation of regression coefficients has great uncertainty. Therefore,

the accuracy of traditional methods based on the estimation of regression coefficients will be greatly unreliable.

2) Significance Test and Significance Factor

Instead of focusing on the least squares estimation of regression coefficients, the SR algorithm proposed in [25] identifies the consumers-branch connectivity based on the significance test of regression coefficients.

The observations of the dependent variable are significantly influenced by the independent variables if the corresponding LNs are connected to this SN. This means that the estimated value of corresponding regression coefficients should be significantly different from 0, and approaching 1. It is equivalent to testing whether hypothesis (20) is rejected, i.e.:

$$\mathrm{H}_0 : \beta_j = 0,\ j \in C \tag{20}$$

According to (15), when the hypothesis (20) is accepted, then there is:

$$\frac{\beta_j^*}{\sigma\sqrt{c_{jj}}} \sim N(0,1) \tag{21}$$

In linear regression, the *t*-statistic is the test statistic for the analysis of variance approach to test the significance of the components in the model. The *t*-statistic can be calculated as:

$$t_{j\text{-}0} = \frac{\beta_j^*}{\sigma^*\sqrt{c_{jj}}} \sim t_{T\text{-}N} \tag{22}$$

where $t_{T\text{-}N}$ is the t distribution with S-N degrees of freedom. Then the probability that the hypothesis (20) holds is:

$$P_0 = P(\beta_j = 0) = P(t_{T\text{-}N} > |t_{j\text{-}0}|) \tag{23}$$

where P_0 represents the P-value for the t-test for $\beta_j = 0$. Similarly, P_1, the P-value for the t-test for $\beta_j = 1$, can be calculated by the above method. So, the smaller the values of P_0 and P_1, the lower the probabilities of $\beta_j = 0$ and $\beta_j = 1$.

According to the linear correlation principle, for the consumer with small error and heavy load, the expected value of the regression coefficient estimation should be close to 1 and the variance should be close to 0. It means that the probability of $\beta_j = 1$ is large and the probability of $\beta_j = 0$ is small. Based on this principle, the significance factor ξ can be defined as:

$$\xi = \ln(P_1/P_0) \tag{24}$$

The ξ values are in the range $(-\infty, +\infty)$. If $P_1 > P_0$, then $\xi > 0$. In extreme cases, if $P_1 = 1$ and $P_0 = 0$, the likelihood for $\beta_j = 1$ will be the highest. Conversely, if $P_1 < P_0$, then $\xi < 0$. In extreme cases, if $P_1 = 0$ and $P_0 = 1$, the likelihood for $\beta_j = 0$ will be the highest. If $P_1 = P_0$ then $\xi = 0$. This means that the likelihood for $\beta_{\Phi j} = 0$ or 1 will be the same at the highest uncertainty.

B. LSR Algorithm

1) SR Algorithm

The SR algorithm is based on the results of the significance test of the independent variables to find a subset of independent variables that explain the highest degree of observations of the dependent variable by iteration. The main approach is that the variables are introduced to the linear regression model in turn, and if the jth independent variable x_j meets the introduction criteria based on its significance, i.e., $P_{0j} = \lambda_{\text{entry}}$, this new variable is introduced. Each time a new variable is introduced, the old variables of the selected equations are tested one by one. If the non-significant exclusion condition is met, i.e., $P_{0i} = \lambda_{\text{remove}}$, then the ith independent variable x_i is removed to ensure that all variables in the subset X_Φ are all significant. This process is repeated several times until no new variable can be introduced. The overall flowchart for the SR algorithm process is shown in Fig. 2. X_Φ is the subset of significant independent variables, X_{remain} is the subset of remaining independent variables, while λ_{entry} and λ_{remove} are the significance thresholds set in advance.

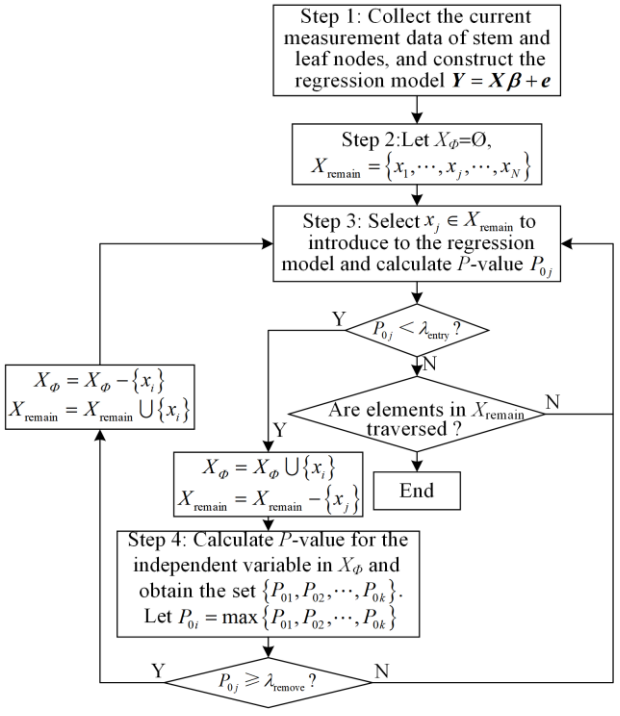


Fig. 2. Flowchart for the SR algorithm process.

In order to avoid getting stuck in an "introduce-remove-introduce" cycle for the same variable, it is generally required that the significance threshold of the introduced independent variable is less than the significance threshold of the removed independent variable, i.e., $\lambda_{\text{entry}} < \lambda_{\text{remove}}$.

The SR method can be applied to find the subset of independent variables which make significant contri-

butions to the dependent variable in the regression model (5). X_Φ reflects the relationship of connectivity between the Φth SN and LNs, which can be called the stem-leaf node subset. The subset of significance factor ξ_Φ is obtained by calculating the significance factors for each independent variable in X_Φ, as:

$$X_\Phi = \{x_\Phi(1),\cdots,x_\Phi(j),\cdots,x_\Phi(n_\Phi)\},\ \forall \Phi \in J,\ \forall j \in C_\Phi \quad (25)$$

$$\xi_\Phi = \{\xi_\Phi(1),\cdots,\xi_\Phi(j),\cdots,\xi_\Phi(n_\Phi)\},\ \forall \Phi \in J,\ \forall j \in C_\Phi \quad (26)$$

2) SR Results in Correction

Each stem-leaf node subset can be obtained according to the SR results, but the apparent and hidden errors can lead to some identification errors or intersections between each subset. Because of the physical limitations, LN cannot connect to different SNs at the same time, so the results should be modified according to the significance factor. Here we define $X_\#$ as the intersection of all stem-leaf node subsets, i.e.:

$$X_\# = X_1 \bigcap X_2 \bigcap \cdots \bigcap X_M \quad (27)$$

$$X_\# = \{x_\#(1),\cdots,x_\#(j),\cdots,x_\#(n_\#)\},\ \forall j \in C_\# \quad (28)$$

$$\xi_{\#\max} = \{\xi_\#(1),\cdots,\xi_\#(j),\cdots,\xi_\#(n_\#)\},\ \forall j \in C_\# \quad (29)$$

$$\xi_\#(j) = \max\{\xi \mid \xi_\Phi(k), x_\Phi(k) = x_\#(j), \Phi \in J, k \in C_\Phi, j \in C_\#\} \quad (30)$$

where $x_\#(j)$ is the LN corresponding to the jth independent variable in intersection $X_\#$; $\xi_{\#\max}$ is the set of maximum values of the significance factors related to the independent variables in intersection $X_\#$; and $\xi_\#(j)$ is the maximum significance factor for $x_\#(j)$ in different X_Φ.

To improve the accuracy of the identification results of the SR algorithm, the identification results should be corrected according to ξ_Φ and $\xi_{\#\max}$, based on the following two rules:

1) For any $x_\Phi(j) \in X_\Phi$, if $\xi_\Phi(j) < 0$, it indicates that the LN has low confidence in belonging to the Φth SN and should be removed from the stem-leaf node subset X_Φ.

2) For any $x_\Phi(j) \in X_\#$, if $\xi_\Phi(j) = \xi_\#(j)$, this indicates that the LN belongs to the Φth SN with the highest confidence and should be removed from the other subset of stem-leaf nodes.

Based on the above rules, the corrected stem-leaf node subset $X_{\mathrm{R}\Phi}$ is obtained as:

$$X_{\mathrm{R}\Phi} = \{x_{\mathrm{R}\Phi}(1),\cdots,x_{\mathrm{R}\Phi}(j),\cdots,x_{\mathrm{R}\Phi}(n_{\mathrm{R}\Phi})\},\ \forall \Phi \in J, \forall j \in C_{\mathrm{R}\Phi} \quad (31)$$

where $x_{\mathrm{R}\Phi}(j)$ is the LN corresponding to the jth significant independent variable in the correct subset $X_{\mathrm{R}\Phi}$.

3) LSR Algorithm Process

The identification results of the SR algorithm can be affected by setting the significance threshold parameter simultaneously. When the significance threshold is set more strictly, LNs with small loads and negligible fluctuations have a high probability of not being selected in a single SR calculation in a subset of stem-leaf nodes, i.e., the stem-leaf connection relationship of this consumer cannot be determined. Thus, this paper improves the SR algorithm and proposes an LSR algorithm for the LVTI problem. The significance threshold is increased and the regression model is updated by a layer-by-layer iterative process in the LSR algorithm to ensure that the topology information of as many consumers as possible is reliably determined. The overall flowchart for the LSR algorithm process is shown in Fig. 3.

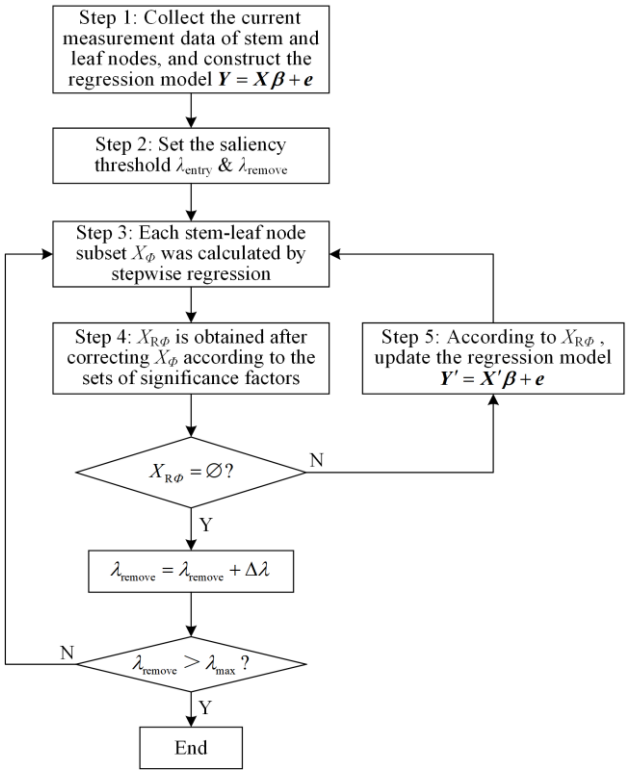


Fig. 3. Flowchart for the LSR algorithm process.

In step 5 of the above algorithm process, $\boldsymbol{Y'}$ is the vector of observations of the dependent variable, calculated as:

$$\boldsymbol{Y'} = \boldsymbol{I}_\Phi - \sum_{j \in X_{\mathrm{R}\Phi}} \boldsymbol{I}_{\mathrm{D}j},\ \forall \Phi \in J \quad (32)$$

where $\boldsymbol{I}_\Phi$ is the vector of injection current magnitude for the Φth stem node; $\boldsymbol{I}_{\mathrm{D}j}$ is the vector of injection current magnitude for the jth leaf node.

$\boldsymbol{X'}$ is the updated design matrix obtained from the specified stem-leaf node dependence, calculated as:

$$\boldsymbol{X'} = [I_{\mathrm{D}ij}]_{T\times N},\ \forall i \in H,\ \forall j \in (C - C_{\Phi\mathrm{all}}) \quad (33)$$

where $C_{\Phi\mathrm{all}}$ is the set of LNs for which the stem-leaf connection relationship is specified and is calculated as:

$$C_{\Phi\mathrm{all}} = X_{\mathrm{R}1} \bigcap \ldots \bigcap X_{\mathrm{R}\Phi} \bigcap \cdots \bigcap X_{\mathrm{R}M},\ \forall \Phi \in J \quad (34)$$

In the LSR algorithm process, when the stem-leaf

connection relationship of the remaining LNs cannot be determined by the SR algorithm, i.e., $X_{R\Phi}=\varnothing$, the significance threshold needs to be increased to ensure that the topology information of more LNs can be determined. If the threshold is set for a significant increase in $\Delta\lambda$, when the increasement reaches a certain level, i.e., $\lambda_{\text{remove}}>\lambda_{\max}$, the layer-by-layer iterative process is suspended and the final identification result is the output. For the remaining and undetermined LNs, the voltage correlation analysis or site survey methods can be considered to identify their stem-leaf connection relationship.

C. Algorithm Performance Evaluation

In order to fully evaluate the performance of the algorithm [25], two metrics are defined: precision $\boldsymbol{\Omega}_{\text{p}}$ and recall $\boldsymbol{\Omega}_{\text{r}}$, which are calculated as:

$$\boldsymbol{\Omega}_{\text{p}}=N_{\text{correct}}/N_{\text{ouput}}\times 100\% \tag{35}$$

$$\boldsymbol{\Omega}_{\text{r}}=N_{\text{correct}}/N\times 100\% \tag{36}$$

where N_{ouput} is the total number of LNs whose stem-leaf connection relationship can be determined by the LSR algorithm; while N_{correct} is the total number of LNs whose stem-leaf connection relationship is correctly determined in the output.

V. Simulation Results

A. Test System

Referring to the simulation of Case 4 in [15], a LVDN test system with a total number of 63 subscribers is established, as shown in Fig. 4.

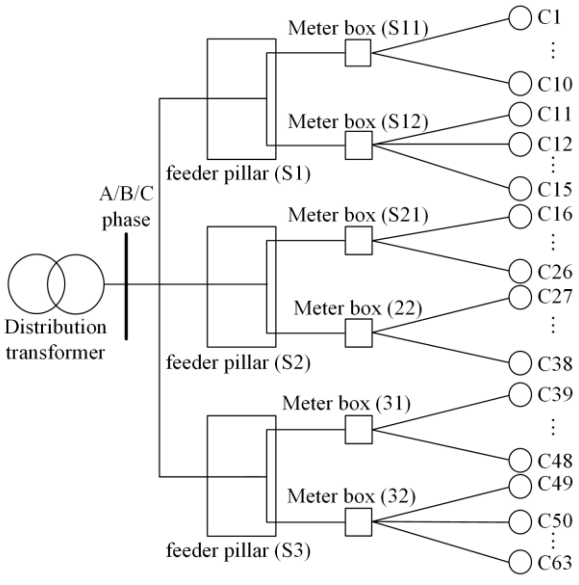


Fig. 4. LVDN test system with 63 users.

As seen, the primary SNs are numbered from S1 to S3, the secondary SNs are numbered from S11 to S32, and the LNs are numbered from C1 to C63.

B. Identification Procedure

The proposed method is validated using real load data from an area in Guangdong, China as an example. Consumer load data are collected at 15-min intervals, and there are a total of 192 time segments. The average value and size of the standard deviation of the active power load during this time period are shown in Fig. 5. The errorless current data in each SN and all LNs are calculated by running power flow 192 times based on the collected load data. The measurement error ($\varepsilon_{\text{s}}=8\%$) is added to the current data in each SN and all LNs to simulate the meter reading errors and clock synchronization errors.

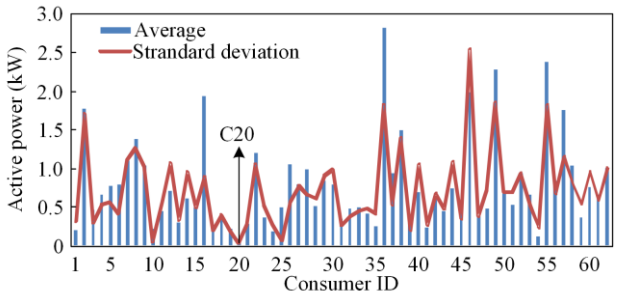


Fig. 5. Mean and standard deviation of consumer load.

The significance threshold is set as $\lambda_{\text{remove}}=0.01$, $\lambda_{\text{entry}}=0.005$, $\Delta\lambda=0.1$, and $\lambda_{\max}=0.5$. Using the injection current of the secondary SNs as the dependent variables and the load current of each LN as the independent variables, the SR algorithm proposed in [25] is used to calculate the stem-leaf node connection relationship. The results are shown in Table I, where the red numbers indicate users with incorrectly identified stem-leaf node connection relationship.

TABLE I
Identification Result of the SR Algorithm

SNs	Subset of LNs
S11	2, 3, 4, 5, 6, 7, 8, 9, 10
S12	10, 11, 12, 13, 14, 15, 50
S21	16, 17, 18, 22, 23, 24, 26
S22	27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38
S31	39, 40, 41, 42, 43, 44, 45, 46, 47, 48
S32	36, 37, 49, 50, 51, 52, 53, 54, 56, 57, 58, 59, 61, 63
Not identified	1, 19, 20, 21, 25, 32, 55, 60, 62

As can be seen from Table I, when the measurement error is large, relying only on the significance test values of P_0 to filter the variables, leads to not only false identification ($\boldsymbol{\Omega}_{\text{p}}=94.4\%$), but also a recall rate $\boldsymbol{\Omega}_{\text{r}}$ of only 85.7%, in addition to the intersection between the subsets of LNs corresponding to different SNs.

To improve the precision rate and avoid the violation of physical constraints due to the intersection between each stem-leaf node subset, the SR results need to be

corrected according to the significance factor. For users of C36, C37, and C50 intersections, the comparison of the significant factor values when they belong to different SNs is shown in Table II.

As can be seen from Table II, although users C36, C37 and C50 are all significant for different SNs (P_0<0.01), the confidence level of intersection users belonging to different SNs can be clearly distinguished based on the significance factor values in Table II. Then, the SR results are corrected by the significance factor to obtain Table III.

TABLE II
SIGNIFICANCE FACTORS FOR C10, C36, C37, C50

LNs	SNs	P_0	P_1	ξ
C10	S11	1.04×10^{-17}	0.97	**39.07**
	S12	0.00341	4.12×10^{-55}	−119.55
C36	S22	3.1×10^{-8}	0.99	**17.27**
	S32	3.3×10^{-3}	6.0×10^{-8}	−10.91
C37	S22	1.2×10^{-44}	0.86	**100.94**
	S32	9.3×10^{-4}	3.89×10^{-8}	−10.08
C50	S12	4.7×10^{-4}	3.9×10^{-119}	−264.9
	S32	4.4×10^{-36}	0.68	**81.01**

TABLE III
SR IDENTIFICATION CORRECTION RESULT

SNs	Subset of LNs
S11	2, 3, 4, 5, 6, 7, 8, 9, 10
S12	11, 12, 13, 14, 15
S21	16, 17, 18, 22, 23, 24, 26
S22	27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38
S31	39, 40, 41, 42, 43, 44, 45, 46, 47, 48
S32	49, 50, 51, 52, 53, 54, 56, 57, 58, 59, 61, 63
Not identified	1, 19, 20, 21, 25, 32, 55, 60, 62

As can be seen from the Table III, the corrected results avoid misidentification and the precision rate can reach 100%. However, because the significance threshold is set conservatively, the recall rate is still only 85.7%. To further improve the recall rate, the LSR algorithm is used and the final identification results are shown in Table IV. The changes in the significance threshold λ_{remove} and the precision and recall rates of the algorithm during the iteration of the LSR algorithm are shown in Fig. 6.

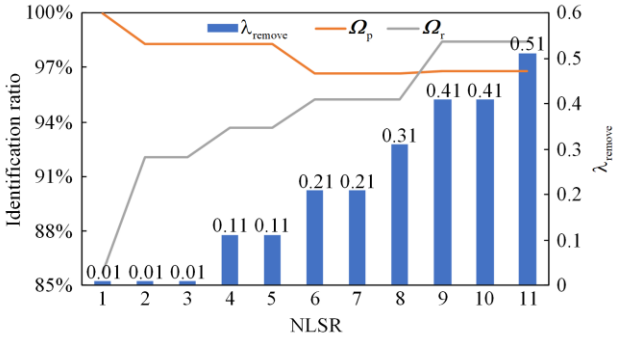


Fig. 6. SR identification correction results of the algorithm under layer-by-layer iteration.

As can be seen from Table IV, although the final identification result has a reduced precision rate (Ω_p = 96.8%), the recall rate is significantly increased (Ω_r = 98.4%), while only the stem-leaf dependence of C20 is not detected. From Fig. 4, the average load and fluctuation of the user C20 are both very small, so the algorithm is unable to determine the stem-leaf node connection relationship.

TABLE IV
IDENTIFICATION RESULTS OF LSR

SNs	Subset of LNs
S11	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
S12	11, 12, 13, 14, 15
S21	16, 17, 18, 19, 21, 22, 23, 24, 26
S22	25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38
S31	39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 55
S32	49, 50, 51, 52, 53, 54, 56, 57, 58, 59, 60, 61, 62, 63
Not identified	20

As can be seen from Fig. 6, the continuous increase of the significance threshold λ_{remove} during the layer-by-layer iterations allows the identification of more stem-leaf dependencies. Thus, the recall rate continues increasing, though the precision rate decreases slightly. In general, the operator can set the threshold $\lambda_{\max}$ according to target preference. The higher the threshold $\lambda_{\max}$, the higher the recall rate, and vice versa.

C. Analysis of the Impact of Hidden Errors

When considering the effect of different types of hidden errors, different procedures M1 (least squares (LS) [31]), M2 (integer quadratic programming (IQP) [16]), M3 (least absolute shrinkage and selection operator lasso regression [24]), and M4 (LSR proposed in this paper) are set and their identification results are compared. There are several hidden error scenarios, which are discussed below.

1) Only a Single Type of Hidden Error Exists

There is only a single category of hidden error such as ETEs, PCEs or CMEs, and the hidden error is added to the measurement data. For example, when there is only ETEs in the LVDN, the current measurements of partial users will be reduced to one-tenth of the true measurements. The precision rates of methods M1 to M4, as well as the recall rates of method M4 are calculated when the hidden error rate ε_{hq}, ε_{hz} or ε_{hk} varies between 0 and 10%, respectively. The results are shown in Fig. 7, where the red curves indicate the recall rate of the M4 method at different error rates, and the different columns indicate the precision rates of methods M1 to M4.

As can be seen from Fig. 7, for the same error size, CMEs can cause more significant decrease in precision rate than the other two types of hidden errors. This is because CMEs can cause significant fitting errors at some time instants, which in turn amplifies SSE and leads to greater uncertainty in the estimated values of

regression coefficients. Therefore, the performances of M1–M3 based on comparison of estimated values of regression coefficients are poor.

The precision rate of the M4 method decreases with the increase of the error rate, but remains above 90%, so the accuracy of the LVTI results can be effectively guaranteed. In addition, M2 performs better than other methods in the absence of hidden error, which indicates that IQP has better performance when only the measurement error is considered.

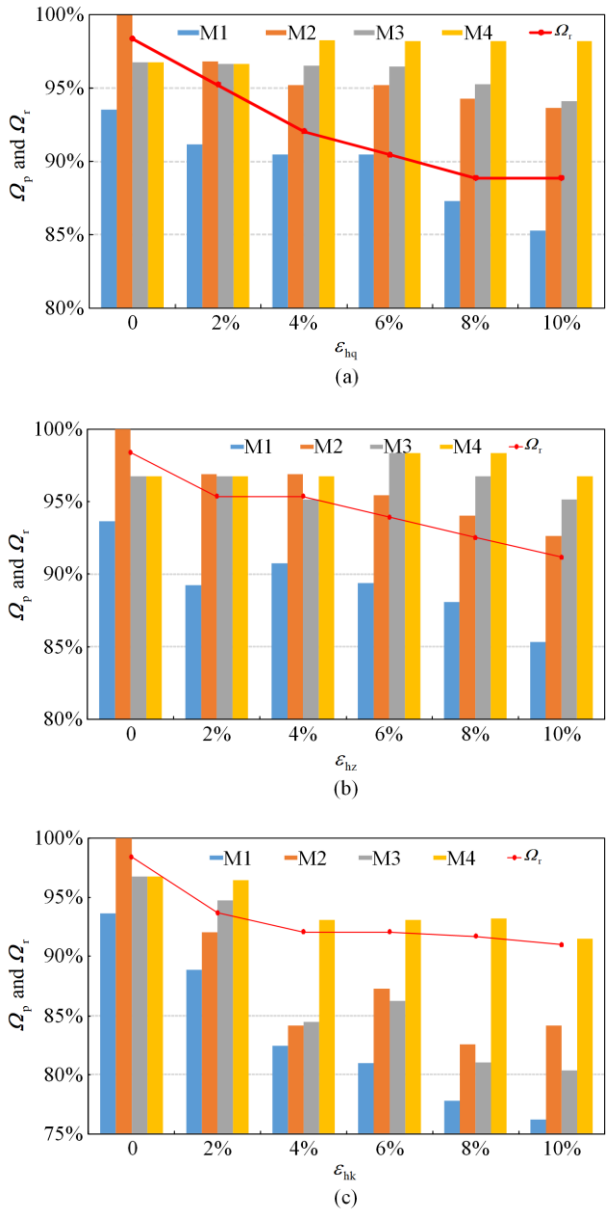


Fig. 7. Identification results for M1 to M4 under the influence of a single type of hidden errors. (a) ETEs. (b) PCEs. (c) CMEs.

2) Multiple Types of Hidden Errors Exist Simultaneously

If multiple types of hidden errors such as ETEs, PCEs, and CMEs exist in the LVDN at the same time, the precision rates of methods M1 to M4, as well as the recall rates of the method M4 are calculated and shown by the red curve, when the hidden error rate ε_h varies from 0 to 25%, in Fig. 8.

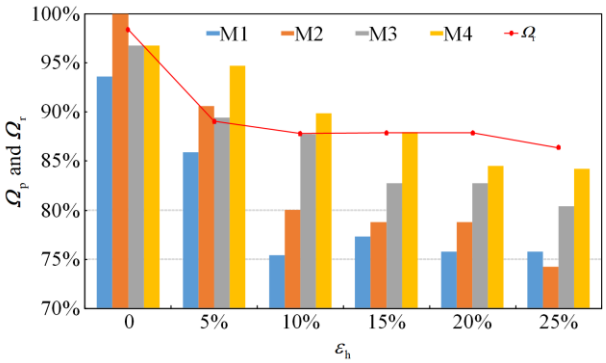


Fig. 8. Identification results for M1 to M4 under combined errors.

As can be seen from Fig. 8, when there are several types of hidden errors at the same time, the precision rates of methods M1 to M4 are lower than those with only ETEs or PCEs, but are slightly better than the CMEs alone. Therefore, the accuracy of the algorithm depends on the type of hidden errors The greatest impact on the identification performance is when considering several types of hidden errors simultaneously. When the error increases further, the precision rate of the proposed method will decrease further while the recall rate maintains at a certain level. This is because the recall rate is mainly determined by the threshold λ_{max}, while the precision rate is influenced by the size of the hidden errors.

There are typically various hidden errors in measurement data in LVDN, and it is impossible to distinguish between the types of hidden error or the proportion of each type. This is the characteristic of hidden error, and is also a problem that most studies tend to overlook. In pilot application, the influence of uncertain hidden errors on the performance of the method can be alleviated by increasing the number of observed samples or removing obviously abnormal sample data.

3) Identification of Stem-leaf Node Connection Relationship at Different Levels

In order to compare the effectiveness of the proposed M4 algorithm for identifying stem-leaf node dependence at different levels, the following two cases are compared.

Case 1: The injected current in the secondary SNs (S11 to S32) is selected as the dependent variables and the outflow current of all LNs is selected as the independent variables.

Case 2: The injected current in primary SNs (S1 to S3) is chosen as the dependent variables and the outflow current of all LNs is chosen as the independent variables.

The precision rates of methods M1 to M4, as well as the recall rate of the method M4 in Case 2 are calculated and shown in Fig. 9.

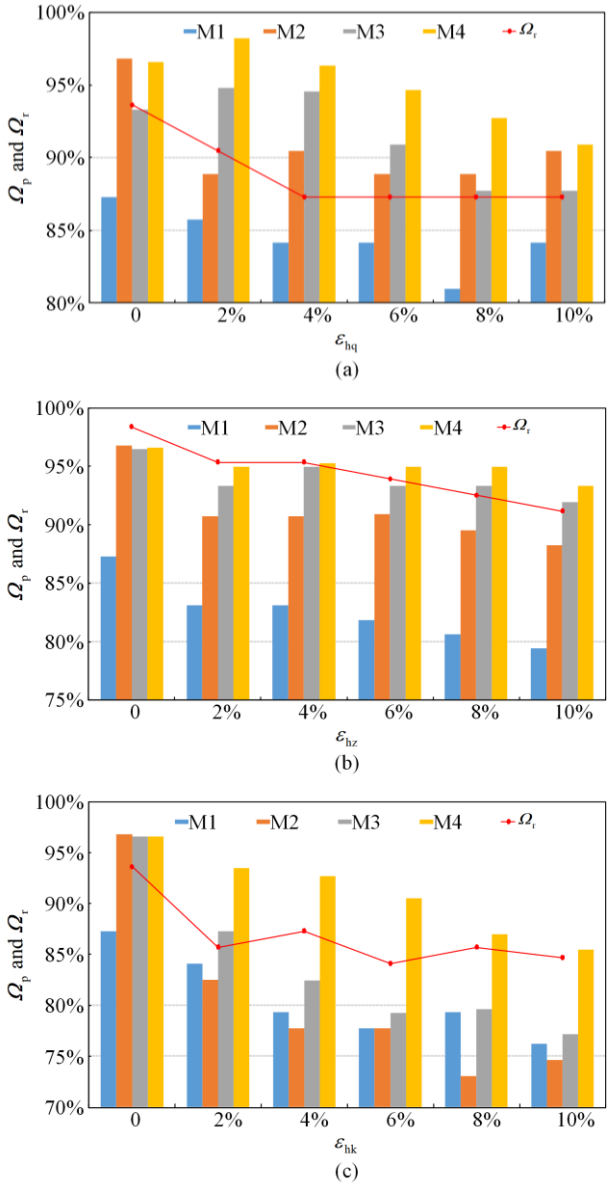


Fig. 9. Identification results for M1 to M4 under the influence of a single type of hidden error in Case 2. (a) ETEs. (b) PCEs. (c) CMEs.

As can be seen from Figs. 8 and 9, when the size of stem-leaf node subsets to be identified is small in Case 2, a large regression fitting error can lead to an increase in the uncertainty of the estimated regression coefficients for the same size of hidden errors. Therefore, the precisions of M1 to M4 in Case 1 are better than those of Case 2 under the influence of a single type of hidden error.

If multiple types of hidden errors such as ETEs, PCEs, and CMEs exist simultaneously, the precision and recall rates are calculated when the error rate ε_h varies from 0 to 25% as shown in Fig. 10.

As can be seen from Fig. 10, the precision and recall rates of the algorithm in Case 1 are better than those of Case 2. It shows that the finer the granularity of the topology identification, the greater the difference in the significance test of regression coefficients. Therefore, better algorithm performance is obtained in Case 2.

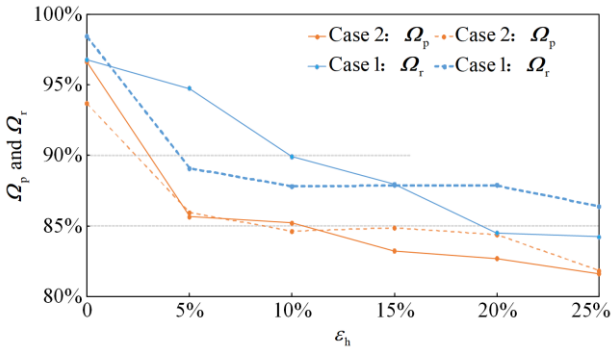
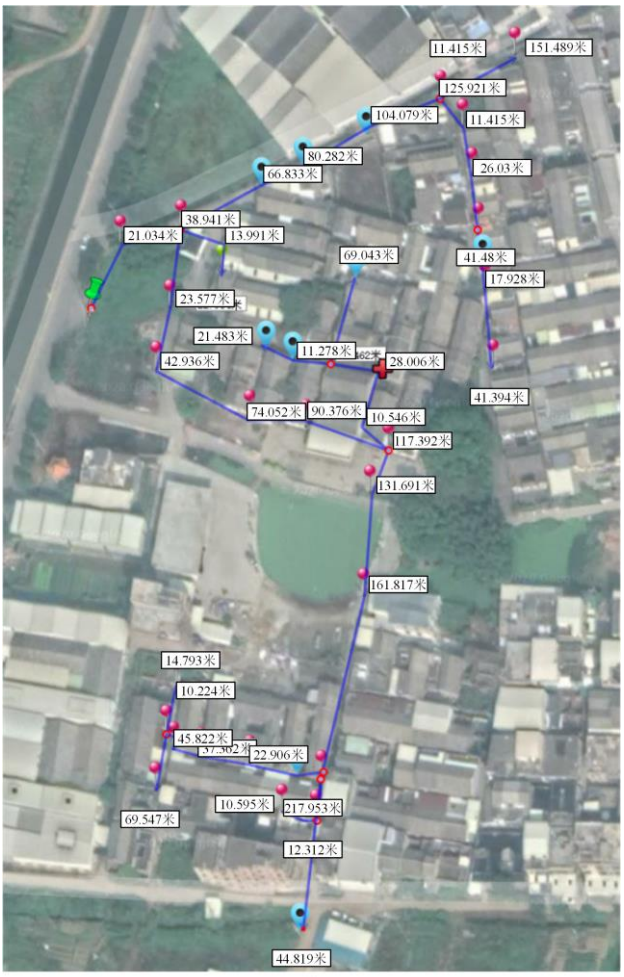


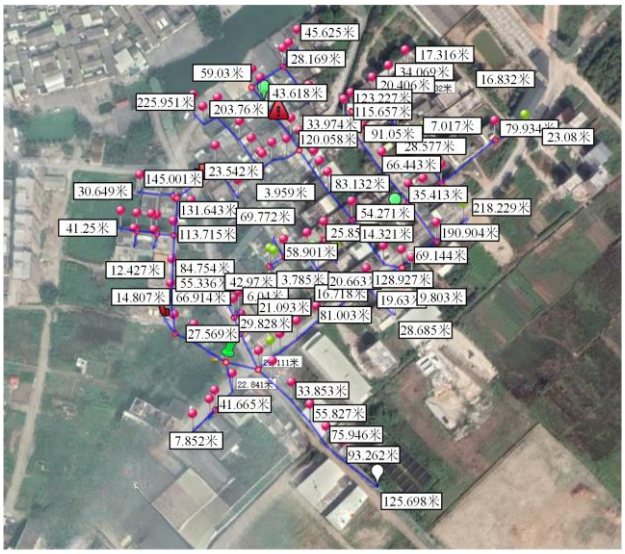
Fig. 10. Identification results for M4 at different levels.

D. Pilot Application

In this paper, four actual LVDN segments in Guangdong, China, are selected as pilot applications. The current and voltage measurement data of each meter in the LVDN are collected at 15-min intervals through the temporary installation of additional communication terminals. Because of the variation in the duration of temporary installation of the terminals, the size of the data samples collected also varies. The geographical location information and topological connectivity of the four LVDN segments are shown in Fig. 11.



(a)



(b)

(c)

Fig. 11. Geographic location information and topological connectivity of four pilot LVDNs. (a) Segment 1. (b) Segment 2. (c) Segment 3 and Segment 4.

The experimental setup of the pilot applications is as follows.

1) Appropriate LVDN segments are chosen as the test samples, and additional communication terminals are installed to collect the measurement data of each meter.

2) The collected measurement data are preprocessed, including removing outliers and filling in missing data.

3) The stem-leaf connection relationship to be identified is determined, and the LSR algorithm proposed in this paper is used to obtain the LVTI results.

4) The performance of the LSR algorithm is evaluated by comparing with the real topology results obtained from manual surveys.

The basic conditions of the four selected pilot LVDN segments are shown in Table V, and the identification results are shown in Table VI.

As can be seen from Table VI, the LSR algorithm proposed in this paper can achieve a precision rate of over 90% when applied to the actual LVDN segment, although the recall rate does not reach 100%, which is mainly because of some customers having light loads or hidden errors, resulting in negligible load characteristics.

TABLE V
PILOT LVDN SEGMENTS

LVDN segments	Number of LNs	Number of time instants	Number of SNs
1	127	576	12
2	141	672	9
3	165	424	3
4	139	288	3

TABLE VI
APPLICATION RESULTS OF THE PROPOSED METHOD

LVDN segment	Ω_p (%)	Ω_r (%)
1	94.5	90.4
2	96.2	92.2
3	91.1	88.1
4	90.3	87.7

VI. CONCLUSION

In this paper, an LSR-based LVTI method is proposed considering hidden errors. The following conclusions are obtained through the simulation analysis of the test system and practical application.

1) The hidden error is a key factor affecting the accuracy of LVTI. The LSR algorithm proposed in this paper can still achieve a precision rate of over 90% when the hidden error rate is less than 10%.

2) Compared with the SR algorithm, the LSR algorithm proposed in this paper can maximize the recall rate by iteratively updating the SR model and significance threshold parameters.

3) Different types of hidden errors affect the accuracy of the LVTI method to different degrees, whereas CMEs have the highest impact.

4) The performance of the proposed method is associated with the granularity of the LVTI, and the finer the granularity of the topology identification, the better the performance.

For future practical applications of the proposed method, the impact of larger hidden errors on the precision rate of the proposed method will be studied. In addition, many state-of-the-art methods, such as generative adversarial networks [32], reinforcement learning [33] also have the potential to achieve better identification performance considering the influences of hidden errors.

ACKNOWLEDGMENT

Not applicable.

AUTHORS' CONTRIBUTIONS

Yongjun Zhang: investigation, methodology, and software. Yingqi Yi: writing original draft. Wenyang Deng: supervision, writing review and editing. Siliang

Liu: writing review and editing. Lai Zhou: data curation. Kaidong Lin: writing review and editing. Yongzhi Cai: funding acquisition. All authors read and approved the final manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China (No. 52177085) and Science and Technology Planning Project of Guangzhou (No. 202102021208).

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

DECLARATIONS

Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

AUTHORS' INFORMATION

Yongjun Zhang received the Ph.D. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2004. Currently, he is a professor with the School of Electric Power, South China University of Technology. His main research interests include reactive power optimization, smart energy, and high-voltage direct current (HVDC) transmission.

Yingqi Yi received the M.S. degree in electrical engineering from South China University of Technology, Guangzhou, China, where he is currently pursuing the Ph.D. degree. His main research interests include the data analytics in distribution system, and optimal operation of distribution network.

Wenyang Deng received the M.S. degree in electrical power system engineering from the University of Manchester, Manchester, U.K., and the Ph. D. degree in electrical computer engineering from the University of Macau, Macau, China. Dr. Deng is now a postdoctoral researcher at the South China University of Technology, Guangzhou, China. His main research interests include the regulation of inverters, power sharing and power quality improvement in microgrids.

Siliang Liu received the M.S. degree in electrical engineering from South China University of Technology, Guangzhou, China. He is now a Ph.D. candidate at the same University. His main research interests include advanced metering infrastructure, and optimal operation of distribution network.

Lai Zhou received the M.S. and Ph.D degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2016 and 2021, respectively. She is currently a lecturer in Guangzhou Panyu vocational and technical college, Guangzhou, China. Her research includes analyzing and controlling the operation of home energy management system, application and analysis of big data in low voltage distribution network, and new energy vehicle battery energy control.

Kaidong Lin received the M.S. degree in Guangdong University of Technology, Guangzhou, China. He is now a Ph.D. candidate at the same University. His main research interests include advanced metering infrastructure, and optimal operation of distribution network.

Yongzhi Cai received the M.S. and Ph.D. degrees in electrical engineering from the South China University of Technology, Guangzhou, China, in 2010 and 2016, respectively. He is currently a senior engineer with Metrology Center Guangdong Power Company Grid, Qingyuan, China. His main research interests include advanced metering infrastructure, and optimal operation of distribution network.

REFERENCES

[1] W. Deng, Y. Zhang, and Y. Tang *et al.*, "A neural network-based adaptive power-sharing strategy for hybrid frame inverters in a microgrid," *Frontiers in Energy Research*, vol. 10, Jan. 2023.

[2] P. Sivalingam and M. Gurusamy, "Momentum search algorithm for analysis of fuel fell vehicle-to-grid system with large-scale buildings," *Protection and Control of Modern Power Systems*, vol. 9, no. 2, pp. 147-160, Mar. 2024.

[3] H. Yin, Y. Xuan, and Y. Huang *et al.*, "Virtual impedance-based low-voltage distribution network topology identification method," *Power System Protection and Control*, vol. 52, no. 3, pp. 83-93, Feb. 2024. (in Chinese)

[4] M. Irving and M. Sterling, "Subsegment data validation," *IEE Proceedings Generation*, *Transmission and Distribution*, vol. 129, no. 3, pp. 119-122, 1982.

[5] B. Xu, G. Zhang, and K. Li *et al.*, "Reactive power optimization of a distribution network with high-penetration of wind and solar renewable energy and electric vehicles," *Protection and Control of Modern Power Systems*, vol. 7, no. 4, pp.1-13, Oct. 2022.

[6] Y. Yi, L. Zhou, and Q. Li *et al.*, "Improving correlation-based consumer phase identification for incomplete data," in *2020 IEEE Sustainable Power and Energy Conference (iSPEC)*, Chengdu, China, Nov. 2020, pp. 2533-2538.

[7] L. Qin, W. Huang, and W. Guo *et al.*, "Topology identification method of low-voltage distribution network based on improved pearson correlation coefficient method," in *2021 IEEE 2nd China International Youth Conference on Electrical Engineering (CIYCEE)*, Chengdu, China, Nov. 2022, pp. 1-6.

[8] T. Short, "Advanced metering for phase identification, transformer identification, and secondary modeling," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 651-658, Jun. 2013.

[9] M. Lave, M. Reno, and J. Peppanen, "Distribution system parameter and topology estimation applied to resolve low-voltage circuits on three real distribution feeders," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1585-1592, Jul. 2019.

[10] V. Cunha, W. Freitas, and F. Trindade *et al*., "Automated determination of topology and line parameters in low voltage systems using smart meters measurements," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5028-5038, Nov. 2020.

[11] Y. Liao, Y. Weng, and G. Liu *et al.*, "Urban MV and LV distribution grid topology estimation via Group Lasso," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 12-27, Jan. 2019.

[12] G. Cavraro, V. Kekatos, and S. Veeramachaneni, "Voltage analytics for power distribution network topology verification," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 1058-1067, Jan. 2019.

[13] C. Lu, L. Zhao, and Y. Li, "Topology checking method for low voltage distribution network based on fuzzy c-means clustering algorithm," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China, Sept. 2020, pp. 1077-1080.

[14] B. Foggo and N. Yu, "A comprehensive evaluation of supervised machine learning for the phase identification problem," *International Journal of Computer and Systems Engineering*, vol. 12, no. 6, pp. 419-427, May 2018.

[15] J. Zhao, L. Li, and Z. Xu *et al.*, "Full-scale distribution system topology identification using markov random field," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 4714-4726, Nov. 2020.

[16] V. Arya, D. Seetharam, and S. Kalyanaraman *et al.*, "Phase identification in smart grids," in *2011 IEEE International Conference on Smart Grid Communications (Smart Grid Comm)*, Brussels, Belgium, Jan. 2011, pp. 25-30.

[17] P. Kumar, V. Arya, and D. A. Bowden *et al.*, "Leveraging DERs to improve the inference of distribution network topology," in *2017 IEEE International Conference on Smart Grid Communications (Smart Grid Comm)*, Dresden, Germany, Jul. 2018, pp. 52-57.

[18] W. Wang, N. Yu, and B. Foggo *et al.*, "Phase identification in electric power distribution systems by clustering of smart meter data," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, CA, USA, Jan. 2016, pp. 259-265.

[19] J. Yu, Y. Weng, and R. Rajagopal, "PaToPa: a data-driven parameter and topology joint estimation framework in distribution grids," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4335-4347, Jul. 2018.

[20] J. Yu, Y. Weng, and R. Rajagopal, "PaToPaEM: a data-driven parameter and topology joint estimation framework for time-varying system in distribution grids," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 1682-1692, May 2019.

[21] M. Xu, R. Li, and F. Li, "Phase identification with incomplete data," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2777-2785, Jul. 2018.

[22] Z. S. Hosseini, A. Khodaei, and A. Paaso, "Machine learning-enabled distribution network phase identification," *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 842-850, Mar. 2021.

[23] J. P. Satya, N. Bhatt, and R. Pasumarthy *et al.*, "Identifying topology of low-voltage distribution networks based on smart meter data," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5113-5122, Sept. 2018.

[24] X. Tang and J. V. Milanovic, "Phase identification of LV distribution network with smart meter data," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*, Portland, OR, Feb. 2019, pp. 1-5.

[25] Y. Yi, S. Liu, and Y. Zhang *et al.*, "Phase identification of low-voltage distribution network based on stepwise regression method," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 4, pp. 1224-1234, Jul. 2023.

[26] C. Si, S. Xu, and C. Wan *et al.*, "Electric load clustering in smart grid: methodologies, applications, and future trends," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 2, pp. 237-252, Mar. 2021.

[27] M. Lisowski, R. Masnicki, and J. Mindykowski, "PLC-enabled low voltage distribution network topology monitoring," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6436-6448, Nov. 2019.

[28] M. Jafarian, A. Soroudi, and A. Keane, "Resilient identification of distribution network topology," *IEEE Transactions on Power Delivery*, vol. 36, no. 4, pp. 2332-2342, Aug. 2021.

[29] B. Appasani, A. V. Jha, and S. K. Mishra *et al*., "Communication infrastructure for situational awareness enhancement in WAMS with optimal PMU placement," *Protection and Control of Modern Power Systems*, vol. 6, no. 1, pp. 1-12, Jan. 2021.

[30] L. Zhou, Q. Li, and Y. Zhang *et al.*, "Consumer phase identification under incomplete data condition with dimensional calibration," *International Journal of Electrical Power & Energy Systems*, vol. 129, Jul. 2021.

[31] F. Wei, Y. Cai, and J. Tang, "Low-voltage station area topology recognition method based on weighted least squares method," in *2020 IEEE Sustainable Power and Energy Conference (iSPEC)*, Chengdu, China, Apr. 2021, pp. 2539-2544.

[32] Q. Sun, C. Ren, and J. Hu *et al.*, "Non-intrusive modeling for integrated energy system based on two-stage GAN," *iEnergy*, vol. 1, no. 2, pp. 257-266, Nov. 2022.

[33] L. Yang, Q. Sun, and N. Zhang *et al*., "Indirect multi-energy transactions of energy internet with deep reinforcement learning approach," *IEEE Transactions on Power Systems*, vol. 37, no. 5, pp. 4067-4077, Sept. 2022.