Phase Identification of Low-voltage Distribution Network Based on Stepwise Regression Method

Yingqi Yi, Siliang Liu, Yongjun Zhang, Ying Xue, Wenyang Deng, and Qinghao Li

***Abstract*—Accurate information for consumer phase connectivity in a low-voltage distribution network (LVDN) is critical for the management of line losses and the quality of customer service. The wide application of smart meters provides the data basis for the phase identification of LVDN. However, the measurement errors, poor communication, and data distortion have significant impacts on the accuracy of phase identification. In order to solve this problem, this paper proposes a phase identification method of LVDN based on stepwise regression (SR) method. First, a multiple linear regression model based on the principle of energy conservation is established for phase identification of LVDN. Second, the SR algorithm is used to identify the consumer phase connectivity. Third, by defining a significance correction factor, the results from the SR algorithm are updated to improve the accuracy of phase identification. Finally, an LVDN test system with 63 consumers is constructed based on the real load. The simulation results prove that the identification accuracy achieved by the proposed method is higher than other phase identification methods under the influence of various errors.**

***Index Terms*—Phase identification, low-voltage distribution network (LVDN), stepwise regression, smart meter, data-driven method.**

NOMENCLATURE

A. Sets

$\zeta_{in\Phi}$	Subset of significant correction factors corresponding to $X_{in\Phi}$
H	Set of indices of measurements
J	Set of indices of phases
L	Set of indices of consumers
$X_{in\Phi}$	Subset of significant independent variables for phase Φ
$X'_{in\Phi}$	Corrected subset of significant independent variable for phase Φ

B. Vectors and Matrices

$\boldsymbol{\beta}$	Vector of regression coefficients
$\boldsymbol{\beta}_{\Phi}$	Vector of regression coefficients for phase Φ
$\boldsymbol{e}$	Vector of errors
$\boldsymbol{e}_s$	Vector of measurement errors
$\boldsymbol{e}_m$	Vector of model errors
$\boldsymbol{e}_h$	Vector of hidden errors
$\tilde{\boldsymbol{I}}_{\Phi}$	Vector of current phasor for phase Φ
$\tilde{\boldsymbol{I}}_{\mathrm{M}}$	Matrix of consumer current phasor
$\boldsymbol{X}$	Design matrix of current magnitudes or active power measurements of consumers in low-voltage distribution network (LVDN)
$\boldsymbol{Y}$	Vector of current magnitudes or active power measurements of phase

C. Variables

$\beta_{\Phi j}$	Regression coefficient of the j^{th} independent variable for phase Φ
β_j	Regression coefficient of the j^{th} independent variable
σ_{si}	Standard deviation of measurement error at the i^{th} time instant
ζ	Significance correction factor of independent variable regression coefficient
Ω_{p}	Precision rate
Ω_{r}	Recall rate
Φ	Index of phases
F_j	F-test value of the j^{th} independent variable
$\tilde{I}_{\Phi i}$	Injection current for phase Φ at the i^{th} time instant
$\tilde{I}_{Mij}$	Load current of the j^{th} consumer at the i^{th} time instant
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
$x_{\Phi(i)}$	Consumer ID corresponding to the i^{th} element in set $X_{in\Phi}$

Manuscript received: October 11, 2022; revised: January 1, 2023; accepted: January 20, 2023. Date of CrossCheck: January 20, 2023. Date of online publication: April 13, 2023.

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

Y. Yi, S. Liu, Y. Zhang (corresponding author), Y. Xue, W. Deng, and Q. Li are with the School of Electric Power, South China University of Technology, Guangdong Key Laboratory of Clean Energy Technology, Guangzhou 510641, China (e-mail: yi_yingqi@foxmail.com; liang_in_ps@163.com; zhangjun@scut.edu.cn; dr.yingxue@foxmail.com; dwyang@scut.edu.cn; liqinhao@scut.edu.cn).

DOI: 10.35833/MPCE.2022.000655

D. Parameters

α	Accuracy level of meter
η	Deviation of meter clock relative to reference clock
ε_s	Measurement error ratio
λ_{entry}	Significance introduced threshold
λ_{remove}	Significance remove threshold
λ_{lasso}	Regularization parameter of Lasso regression
g	Lower limit of allowable recall rate
n_Φ	Total number of elements for subset $X_{in\Phi}$
P_0	*P*-value for the *F*-test that regression coefficient is equal to 0
P_1	*P*-value for the *F*-test that regression coefficient is equal to 1

I. Introduction

In recent years, with the application of information and communication technology (ICT) in power systems, "digitization" has become an important feature of the modern power system. Among them, advanced metering infrastructure (AMI) provides the foundation to transform the planning, operation, and management of distribution networks [1], [2].

The low-voltage distribution network (LVDN) is located at the edge of the power system and directly connects consumers. It is critical to ensuring the quality of the power supply and improving the consumer experience. However, the consumer phase connectivity information in LVDN is generally missing or inaccurate, which has become a bottleneck that restricts the planning and operation management of LVDN [3]. With the rapid application of smart meters in LVDN, a large number of researchers have studied the phase identification of LVDN by analyzing the data from smart meters. These studies can be divided into two categories: one is based on the principle of voltage correlation and the other is based on the principle of energy conservation [4].

The principle of voltage correlation means that the correlation factors between the voltage profiles of consumers reflect the electrical distance between them, and consumers with a close electrical distance will have a greater probability of being in the same phase connectivity or the same branch [5]. References [6] and [7] analyzed the correlation factors between the voltage profiles of consumers and the voltage profiles of each phase and identified the consumer phase connectivity by the correlation factors. Reference [8] identified the consumer phase connectivity by analyzing the correlation between the consumer and the voltage profiles of each phase of the three-phase feeder meter. The above method has poor identification performance when the load is light or the level of three-phase unbalance is low in LVDN. References [9] and [10] used the linear regression method to identify the parallel connections of branches, consumer phase, and line impedance based on the line voltage drop model. Reference [11] extended the application of this method to three-phase four-wire LVDN, making it applicable to the LVDNs in Europe and South America. However, it is difficult to guarantee its performance for LVDNs with more complex structures or with a large number of consumers, which are connected to a medium-voltage/low-voltage (MV/LV) transformer, such as the LVDNs in China. Reference [12] analyzed the applications of various supervised learning algorithms on the phase identification problem. Reference [13] proposed the selection principle of labelled samples in supervised learning based on information loss theory, which improved the efficiency of sample collection and processing. Reference [14] used the *K*-means clustering algorithm to identify the consumer phase based on the voltage data. In addition, the unsupervised learning algorithms such as spectral clustering [15] and fuzzy C-means clustering [16] are also widely used in the topology identification of LVDN. Although the above-mentioned artificial intelligence recognition methods are easy to apply, their performances in practical applications are not ideal due to the poor interpretability of the models [17].

The principle of energy conservation means that the current injected from the upstream nodes (busbars of each phase) of the LVDN at any point in time is equal to the sum of the currents flowing to the downstream nodes (consumers). Reference [18] proposed an integer programming model and its relaxation method for solving the phase identification of LVDN. Reference [19] proposed that the similarity of load curves of different consumers can be reduced by adjusting the output of distributed generation from the consumer side, thereby improving the accuracy of phase identification based on the integer programming model. Reference [20] considered the nonlinear power flow equation constraints and converted the topology identification problem of LVDN into mixed-integer linear program (MILP), which improved the efficiency of the solution. As this method needs to collect the phase angle information, it requires the installation of phaser measurement units (PMUs), where the investment cost can be high [21].

Compared with integer programming methods, regression analysis and machine learning methods have higher accuracy in phase recognition, which has recently attracted researchers' attention. Reference [22] proposed a phase identification method of LVDN based on Lasso regression. Reference [23] and [24] proposed to extract the high-frequency components in the time series of load based on the Fourier transform, and then used the improved clustering method for phase identification. This method can to some extent solve the problem of incomplete data caused by the limited coverage of smart meters. Reference [25] proposed to use the principal component analysis (PCA) method to identify the consumer phase. A synthesis of the advantages and disadvantages of different methods is shown in Table I.

In general, the identification methods of LVDN topology based on energy conservation usually require high-quality measurement data. However, in fact, due to the impacts of meter measurement errors, clock synchronization errors, communication interruptions, and other negative factors, the measurements may be seriously distorted, and the identification accuracy cannot be guaranteed [26].

Therefore, this paper proposes to apply a stepwise regression (SR) algorithm to effectively identify phase connection for consumers in LVDN. SR is a systematic algorithm for

adding and removing terms from a multiple linear model based on their statistical significance in a regression [27]. SR algorithm has been applied in many problems in which the influence of measurement noise cannot be ignored [28], [29].

TABLE I
AVAILABLE LITERATURE ON IDENTIFICATION METHODS OF LVDN TOPOLOGY

Category	Reference	Method	Advantages and disadvantages
Voltage correlation	[6]-[8]	Correlation analysis	(+) Only require voltage data, with high computational efficiency (–) Have poor identification performance on account of short electrical distances, light load, or balance three-phase load
	[9]-[11]	Linear regression	(+) Identify the connections of branches, consumer phase, and line impedance simultaneously (–) Be unavailable for LVDNs with more complex structures or with a large number of consumers
	[12], [13]	Supervised learning	(+) Achieve high accuracies based on sufficient training data samples (–) Be difficult to obtain training data labels in practice
	[14]-[16]	Clustering	(+) Be easy to implement and tune (–) Be sensitive to algorithm parameter
Energy conservation	[18], [19]	Integer programming	(+) Only require current data (–) Be sensitive to bad and incomplete data
	[20]	MILP	(+) Have high computational efficiency (–) Need to collect phase angle information
	[22]	Lasso regression	(+) Achieve higher accuracies based on strict parameter (–) Be sensitive to bad and incomplete data
	[23], [24]	Clustering	(+) Adapt to incomplete data (–) Require high data synchronization
	[25]	PCA	(+) Only require load data and have high computational efficiency (–) Be sensitive to bad and incomplete data

Note: the symbols (+) and (–) represent advantages and disadvantages, respectively.

The main contributions of this paper are as follows.

1) This paper applies the SR algorithm to identify phase connection for consumers in LVDN for the first time. The algorithm identifies the phase connectivity according to the significance test.

2) The significance correction factors are defined in this paper to correct the results of the SR algorithm, which can improve the identification accuracy with a variety of errors.

3) The size of errors when selecting current and active power as regression variables is analyzed, and the effects of selecting different regression variables on the accuracy of the phase identification are compared.

4) The influence of different algorithm parameters on the identification accuracy of the proposed method is analyzed, and on the premise of considering various errors, it is compared with the least square (LS) method [30], integer quadratic programming (IQP) [17], Lasso regression method [22], and voltage correlation comparison [8].

II. PROBLEM DESCRIPTION

Different from the LVDN in North America, LVDN in China generally has a three-phase four-wire structure [31], as shown in Fig. 1. It can be observed from Fig. 1 that each single-phase consumer (e.g., C21, C11) is connected to the A/B/C phase feeder and neutral (N) through the service wire. The sensors installed at the LV side of the distribution transformer can obtain the current or active power data [32] with a 15-min resolution. The proposed data-driven method only needs to use the current or active power data measured by smart meters or sensors.

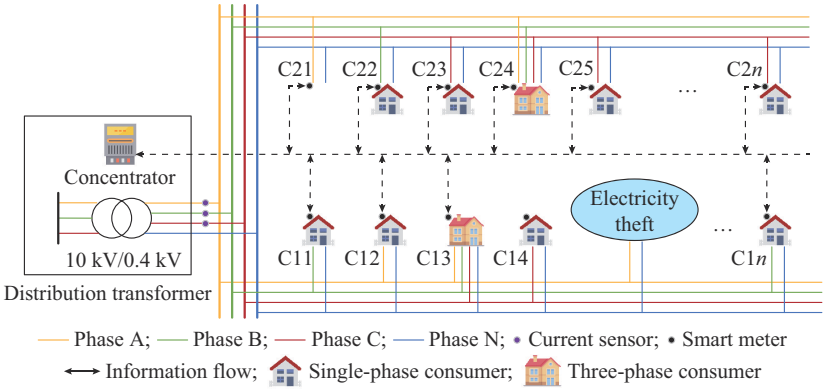

Fig. 1. Illustration of a simple LVDN.

The phase topology of LVDN can be considered to be the connectivity relationship between consumers and each phase feeder. As shown in Fig. 2, the principle of energy conservation implies that the energy of the incoming phase feeders

are equal to the sum of energy of outgoing consumers connected to that phase [25]. This principle leads a set of linear equations. For the mathematical description of the problem, define $J=\{\mathrm{A},\mathrm{B},\mathrm{C}\}$, $H=\{1,2,\ldots,T\}$, $L=\{1,2,\ldots,N\}$, where T is the number of measurements, and N is the number of consumers connected to the MV/LV distribution transformer.

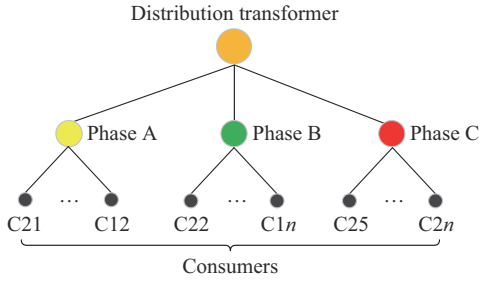

Fig. 2. Phase topology of LVDN.

Then, the vector of phase current phasor $\tilde{\boldsymbol{I}}_{\Phi}\in\mathbf{R}^{T}$, the matrix of consumer current phasor $\tilde{\boldsymbol{I}}_{M}\in\mathbf{R}^{T\times N}$, and the vector of regression coefficient $\boldsymbol{\beta}_{\Phi}\in\mathbf{R}^{N}$ can be expressed as:

$$\tilde{\boldsymbol{I}}_{\Phi}=[\tilde{I}_{\Phi 1}\quad \tilde{I}_{\Phi 2}\quad \ldots\quad \tilde{I}_{\Phi i}\quad \ldots\quad \tilde{I}_{\Phi T}]\quad \forall i\in H,\forall \Phi\in J \tag{1}$$

$$\tilde{\boldsymbol{I}}_{M}=[\tilde{I}_{Mij}]_{T\times N}\quad \forall i\in H,\forall j\in L \tag{2}$$

$$\boldsymbol{\beta}_{\Phi}=[\beta_{\Phi 1}\quad \beta_{\Phi 2}\quad \ldots\quad \beta_{\Phi j}\quad \ldots\quad \beta_{\Phi N}]^{\mathrm{T}}\quad \forall j\in L,\forall \Phi\in J \tag{3}$$

where $\beta_{\Phi j}=\{0,1\}$ represents the phase connectivity information for the j^{th} consumer: 1 means that this consumer is connected to phase Φ, and 0 means that it is not connected to phase Φ. Three-phase consumers can be equivalent to three single-phase consumers. According to Kirchhoff's current law (KCL), these linear equations can be expressed as:

$$\tilde{\boldsymbol{I}}_{\Phi}=\tilde{\boldsymbol{I}}_{M}\boldsymbol{\beta}_{\Phi} \tag{4}$$

The phase identification essentially solves (4) to obtain the regression coefficient vector, which reflects the corresponding phase connection. Considering that the phase angle data cannot be measured by sensors or smart meters in LVDN [21], the current magnitude measurements or active power measurements are used to replace the current phasor measurements in practice. This approximation inevitably introduces the model errors. Thus, (4) can be rewritten by taking into account both the measurement errors and the influence of hidden errors:

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

The phase current or active power is taken as the dependent variable $\boldsymbol{Y}$, and the current or active power obtained by the consumer meter is taken as the independent variable $\boldsymbol{X}$.

In (5), the errors $\boldsymbol{e}\in\mathbf{R}^{T}$ include the measurement errors $\boldsymbol{e}_{s}\in\mathbf{R}^{T}$, model errors $\boldsymbol{e}_{m}\in\mathbf{R}^{T}$, and hidden errors $\boldsymbol{e}_{h}\in\mathbf{R}^{T}$, i.e.,

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

The measurement errors $\boldsymbol{e}_{s}$ are from the meter reading and the clock synchronization. $\boldsymbol{e}_{s}$ can also be modelled to be Gaussian distribution with an expected value of 0 [25]. Its variance σ_{si} is mainly related to the load profile, the metering error ratio ε_1, and the clock synchronization error ratio ε_2. The distribution of the random error e_{si} at the i^{th} time instant is given by:

$$e_{si}\sim N(0,\sigma_{si}^{2}) \tag{7}$$

$$\sigma_{si}=\varepsilon_{s}\sum_{j\in N}X_{Mij} \tag{8}$$

$$\varepsilon_{s}=\varepsilon_{1}+\varepsilon_{2} \tag{9}$$

$$\varepsilon_{1}=\alpha/3 \tag{10}$$

$$\varepsilon_{2}=\eta/45 \tag{11}$$

where α is the accuracy level of the meter, which is generally 0.2, 1, 2, or 5 [33], and the corresponding value of ε_1 then varies from 0.1% to 1.7%; η is the deviation of the meter clock relative to the reference clock, which is generally 0-5 min; the range of ε_2 is 0-6.7% if the sampling time interval of 15 min is used as the reference; ε_s is the measurement error ratio ranging from 0.1% to 8.4% considering metering errors and clock synchronization errors simultaneously; and X_{Mij} is the current magnitude or active power measurement of the j^{th} consumer at the i^{th} time instant.

The model errors $\boldsymbol{e}_{m}$ refer to the errors caused by ignoring the phase angle and technical losses. These errors are mainly related to the grid structure and the network load level.

The hidden errors $\boldsymbol{e}_{h}$ refer to the missing or serious distortion of measurement data due to the problems such as electricity theft [34], PLC crosstalk [35], and interruption of communication [36]. This error is generally difficult to detect so we name it a hidden error. The values of hidden errors vary widely and are determined by the grid communication status and the network operating conditions.

With the access to distributed energy resources (DERs), there are also integrated prosumers. Since the DERs are generally installed behind consumer meters, the meter outputs are the net imbalance of local demand and supply. Thus, (5) inherently considers the integrated prosumers in LVDN [25]. Hence, the law of energy conservation for the access of integrated prosumers also holds.

III. Phase Identification Method of LVDN

To solve (5), the traditional methods used to convert it into an optimization problem [17] and various optimization techniques could be applied to get the optimal mathematical solution of $\boldsymbol{\beta}$ based on the observations over a period of time [24]. Then, the consumer phase can be identified based on the regression coefficient value [30]. However, owing to the influence of errors, especially hidden errors, the measurements could be severely distorted, which could lead to the significant deterioration of identification results. These optimization methods are simple and straightforward, but they cannot accommodate the influence of errors.

This paper proposes to apply an SR algorithm to identify the consumer phase of LVDN according to their significances, which can be checked through F-test. Instead of focusing on the specific value of the regression coefficient obtained through optimization, the SR method is used to solve (5) based on the P-value, which can reflect the significant contribution of the corresponding independent variable to the observations. Therefore, the SR method provides a systematic

way of identifying the consumer phase connectivity in a statistical framework, which considers errors and can achieve a higher identification accuracy. The key steps of SR algorithm are described as follows.

1) Based on the observations of current or active power, the multi-linear model as shown in (5) can be established.

2) The SR algorithm is used to add and remove the independent variables from (5) based on their statistical significances. The consumer phase is then determined according to the significant independent variable subsets for each phase.

3) Based on the correction factor, the calculation results of the SR algorithm are corrected to obtain the final identification of the consumer phase connectivity.

4) For cases where consumer phase connectivity cannot be determined due to light load or error influence, the methods such as voltage correlation analysis or field testing can be adopted.

A. Significance Test

The independent variable significantly influences the observations of the dependent variable when there is a linear correlation independent variable and a dependent variable. This means that the corresponding regression coefficients should be significantly different from 0. From the perspective of hypothesis testing, it is equivalent to testing whether hypothesis (12) is accepted.

$$H_0: \beta_j = 0 \quad \forall j \in L \tag{12}$$

In this paper, the F-test is used to test the significance of the regression coefficient of a single variable, and the F-test value of the j^{th} variable F_j is constructed as:

$$F_j = (T-N-1)\cdot \Delta SSE_j / SSE \tag{13}$$

$$\Delta SSE_j = SSE_j - SSE \tag{14}$$

$$SSE = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{15}$$

where SSE is the residual sum of squares obtained by linear regression (5) of the dependent variable on the N independent variables; SSE_j is the residual sum of squares obtained by linear regression (5) of the dependent variable on the remaining $N-1$ independent variables after removing the j^{th} independent variable; and ΔSSE_j is the partial residual sum of squares, and its value is equal to the difference between SSE_j and SSE.

Assuming regression error satisfies the normal distribution, F_j will obey the F-distribution with degrees of freedom $(1, T-N-1)$ when $\beta_j = 0$, i.e.,

$$F_j \sim F(1, T-N-1) \tag{16}$$

Then, the probability that the hypothesis holds is:

$$P_{0j} = P(\beta_j = 0) = P(F > F_j) \tag{17}$$

where P_{0j} is called the P-value for the F-test when regression coefficient $\beta_j = 0$. A smaller P-value indicates a higher likelihood that the corresponding independent variable has significant contributions to the observation of dependent variable, and vice versa.

B. SR Algorithm

SR is an iterative procedure to find the subset of the independent variable and corresponding regression coefficients that "best" explain the observations of the dependent variable. The main idea of the SR algorithm is to introduce the variables one by one, and if the j^{th} independent variable 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., $P_{0j} > \lambda_{\text{remove}}$, the old variable is removed to ensure that the variables in the independent variable subset are all significant. This process is repeated by several times until no new variables can be introduced.

In order to avoid falling into the infinite loop of introducing-removing-introducing the same variable, it is generally required that λ_{entry} is smaller than λ_{remove}, i.e., $\lambda_{\text{entry}} < \lambda_{\text{remove}}$. The detailed calculation procedure of the SR algorithm is explained below.

Algorithm 1: SR algorithm

Inputs: observation vector $\boldsymbol{Y}$, design matrix $\boldsymbol{X}$, significance thresholds λ_{entry} and λ_{remove}

Outputs: significant independent variable subset $X_{\text{in}\Phi}$

Step 1: start with initial regression model only with the DC component

Step 2: select and add one independent variable x_j to the regression model. The significance is checked using the F-test to obtain the P-value P_{0j}

Step 3: if $P_{0j} < \lambda_{\text{entry}}$, x_j shall be added to the regression model, and $X_{\text{in}\Phi} = X_{\text{in}\Phi} \cup \{x_j\}$. If $P_{0j} \geq \lambda_{\text{entry}}$, go directly to *Step 6*

Step 4: the significance of all independent variables in regression model shall be checked using the F-test to obtain a set of significant P-value $\{P_{01}, P_{02}, \ldots, P_{0k}\}$

Step 5: let $P_{0i} = \max\{P_{01}, P_{02}, \ldots, P_{0k}\}$. If $P_{0i} \geq \lambda_{\text{remove}}$, x_j shall be removed to the regression model, and $X_{\text{in}\Phi} = X_{\text{in}\Phi} - \{x_i\}$. If $P_{0i} < \lambda_{\text{remove}}$, go directly to *Step 6*

Step 6: *Steps 2-5* are repeated until no independent variable needs to be added or removed from the regression model according to F-test

Step 7: End

As explained above, using the SR algorithm, the significant independent variable subset for each phase can be obtained as:

$$X_{\text{in}\Phi} = \{x_{\text{in}\Phi}(1),\ x_{\text{in}\Phi}(2),\ \ldots,\ x_{\text{in}\Phi}(i),\ \ldots,\ x_{\text{in}\Phi}(n_\Phi)\} \quad \forall \Phi \in J \tag{21}$$

Considering the influence of errors and the settings of significance threshold, there could be intersections among X_{inA}, X_{inB}, and X_{inC}, as shown in Fig. 3, indicating a single-phase consumer is connected to multiple phases. However, it is not possible in practice because single-phase consumers cannot be connected to different phases at the same time. So, it is necessary to correct the results directly obtained from SR algorithm.

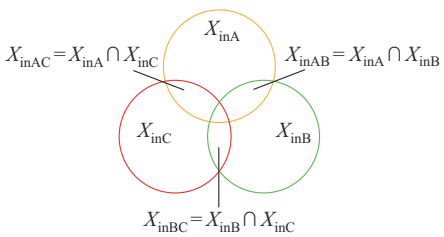

Fig. 3. Intersections among subsets of variables for each phase.

C. Correction of Results from SR Algorithm

According to the linear correlation principle, for the consumers with smaller errors and heavier load, the expected value of the corresponding regression coefficient should be closer to 1 and the variance should be smaller. It means that the likelihood of $\beta_j = 1$ is larger and the likelihood of $\beta_j = 0$ is smaller. Based on this principle, the significance correction factor ζ can be defined as:

$$\zeta = \ln(P_1/P_0) \tag{18}$$

where P_1 represents the P-value for the F-test of $\beta_j = 1$. The smaller the value, the smaller the likelihood of $\beta = 1$. Its calculation method is similar to that for P_0.

The ζ values are in the range of $(-\infty, +\infty)$. If $P_1 > P_0$, $\zeta > 0$. In extreme cases, if $P_1 = 1$ and $P_0 = 0$, this means that the likelihood for $\beta = 0$ will be highest. Conversely, if $P_1 < P_0$, $\zeta < 0$. In extreme cases, if $P_1 = 0$ and $P_0 = 1$, it means that the likelihood for $\beta = 1$ will be highest. If $P_1 = P_0$, $\zeta = 0$. This means that the likelihood for $\beta = 0$ or $\beta = 1$ will be the same at the highest uncertainty.

For $\forall x_i \in X_{\text{in}\Phi}$, if $\zeta_{\text{in}\Phi}(x_i) \le 0$, the reliability of the phase identification result on the corresponding consumer x_i shall be unacceptable. For the intersections of subsets X_{inA}, X_{inB}, and X_{inC}, taking the intersection set X_{inAB} as an example $\forall x_i \in X_{\text{inAB}}$, if $\zeta_{\text{inA}}(x_i) > \zeta_{\text{inB}}(x_i)$, the phase of consumer x_i is more likely to be phase A rather than phase B. In this way, the identification results from the SR algorithm can be corrected based on the values of ζ. The detailed steps are described as follows.

Algorithm 2: correction of results from SR algorithm

Inputs: subset of significant independent variables for phase Φ: $X_{\text{in}\Phi}$

Outputs: corrected subset of significant independent variables for phase Φ $X'_{\text{in}\Phi}$

Step 1: calculate the significant correction factor sets $\zeta_{\text{in}\Phi}$ corresponding to $X_{\text{in}\Phi}$

Step 2: find the elements that are less than 0 in $\zeta_{\text{in}\Phi}$, and remove the corresponding independent variables from $X_{\text{in}\Phi}$

Step 3: for $\forall x_i \in X_{\text{inA}} \bigcap X_{\text{inB}}$, if $\zeta_{\text{inA}}(x_i) < \zeta_{\text{inB}}(x_i)$, remove x_i from X_{inA}; if $\zeta_{\text{inA}}(x_i) > \zeta_{\text{inB}}(x_i)$, remove x_i from X_{inB}; otherwise, remove x_i from both X_{inA} and X_{inB}

Step 4: for $\forall x_i \in X_{\text{inA}} \bigcap X_{\text{inC}}$, if $\zeta_{\text{inA}}(x_i) < \zeta_{\text{inC}}(x_i)$, remove x_i from X_{inA}; if $\zeta_{\text{inA}}(x_i) > \zeta_{\text{inC}}(x_i)$, remove x_i from X_{inC}; otherwise, remove x_i from both X_{inA} and X_{inC}

Step 5: for $\forall x_i \in X_{\text{inB}} \bigcap X_{\text{inC}}$, if $\zeta_{\text{inB}}(x_i) < \zeta_{\text{inC}}(x_i)$, remove x_i from X_{inB}; if $\zeta_{\text{inB}}(x_i) > \zeta_{\text{inC}}(x_i)$, remove x_i from X_{inC}; otherwise, remove x_i from both X_{inB} and X_{inC}

Step 6: repeat *Steps 3-5* to get $X'_{\text{in}\Phi}$ form $X_{\text{in}\Phi}$

Step 7: end

D. Evaluation of Algorithm Performance

The final identification result of the consumer phase connectivity is obtained according to $X'_{\text{in}\Phi}$. In order to evaluate the performance of the algorithm, two indicators, i.e., precision rate Ω_{p} and recall rate Ω_{r}, are proposed as:

$$\Omega_{\text{p}} = (N_{\text{correct}}/N_{\text{output}}) \times 100\% \tag{19}$$

$$\Omega_{\text{r}} = (N_{\text{output}}/N) \times 100\% \tag{20}$$

where N_{output} is the number of consumers with identifiable phase connectivity information from Algorithm 1 and Algorithm 2; and N_{correct} is the number of consumers with correct phase identification from the outputs of the two algorithms.

The output results of the algorithm under different significance thresholds are calculated since it is difficult to obtain the optimal threshold in advance in practical applications. To facilitate the comparison, define the credible precision rate $\Omega_{\text{p,ave}|g}$, which represents the average precision rate under different threshold values when the recall rate Ω_{r} is larger than g:

$$\Omega_{\text{p,ave}|g} = \frac{1}{1-g}\int_g^1 \Omega_{\text{p}}(x)\text{d}x \tag{21}$$

where $\Omega_{\text{p}}(x)$ is the precision rate when $\Omega_{\text{r}} = x$; and g is the lower limit of the allowable recall rate.

IV. Simulation Results

A. Test System

The real LVDN of Guangdong Province in China is used to test the performance of the proposed method. Only the single line diagram of phase A of the test network is shown in Fig. 4 due to the page limit. There are 63 mixed residential or commercial customers in the real LVDN. The consumer IDs of phases A, B, and C are 1-21, 22-42, and 43-63, respectively. The main line model is BLV-150, and the service drop line model is BLV-50. The length of each line is indicated in Fig. 4.

B. Identification Procedure

The consumer load data are collected with a sampling interval of 15 min. The sampling period is 2 days with a total of 192 time instants. The power consumption summary of each phase consumed in 2 days is presented in Fig. 5. The corresponding current or active power on each phase is obtained by executing power flow 192 times. Considering a measurement error ratio $\varepsilon_{\text{s}} = 8\%$, the significance thresholds $\lambda_{\text{remove}} = 0.05$ and $\lambda_{\text{entry}} = \lambda_{\text{remove}}/2 = 0.025$, and the consumer phase identification results are shown in Table II using Algorithm 1. The numbers in bold indicate incorrect phase identification result.

There will be a number of misidentifications ($\Omega_{\text{p}} = 72.4\%$, $\Omega_{\text{r}} = 92.0\%$) with a low precision rate of 72.4%. Also, there are a lot of intersections between the independent variable subsets for each phase in Table II.

To improve the accuracy of identification, the significance correction factor ζ of each variable in $X_{\text{in}\Phi}$ is calculated, as shown in Fig. 6 and Fig. 7.

According to the significance correction factors in Fig. 6 and Fig. 7, the results from Table II are corrected by Algorithm 2. The corrected identification results are shown in Table III, where the numbers in bold indicate incorrect identification.

It can be observed from Table III that the corrected results have a much lower number of misidentifications and the accuracy of the identification is significantly improved to $\Omega_{\text{p}} = 96.5\%$. The recall rate is slightly decreased to 90.4%. Since the output accuracy is of greater importance for engineering applications, it is acceptable to sacrifice a small amount of recall rate to significantly improve the accuracy rate.

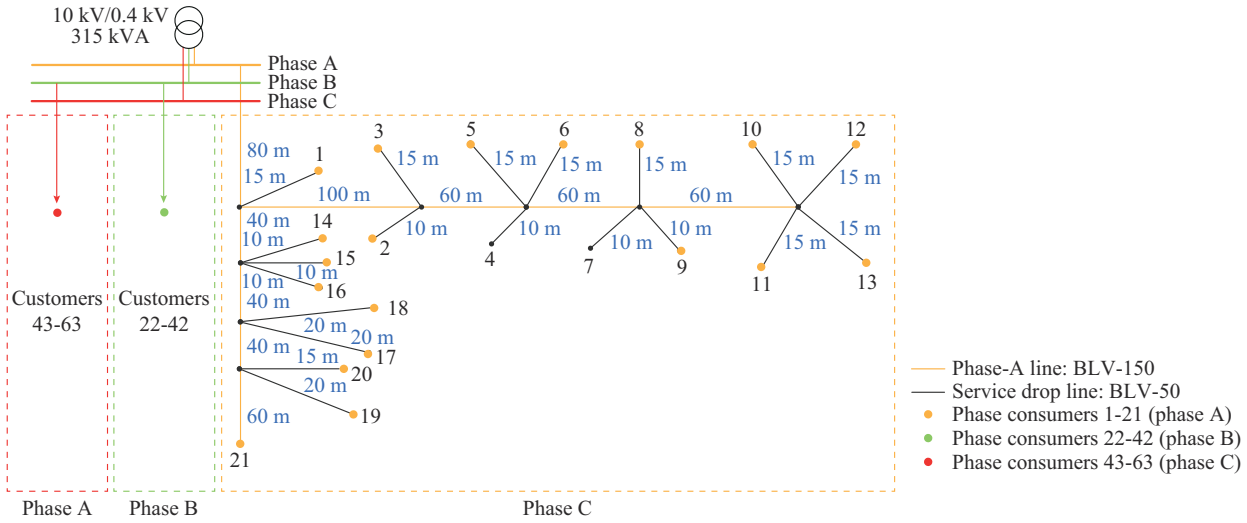

Fig. 4. LVDN test system with 63 consumers.

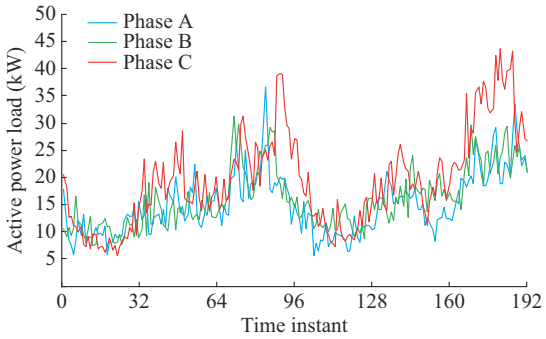

Fig. 5. Power consumption summary of each phase.

TABLE II
IDENTIFICATION RESULTS USING SR ALGORITHM

Subset	Consumer ID
X_{inA}	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 18, 19, **23**, **31**, **38**, **44**, **48**, **53**
X_{inB}	**3**, **10**, **14**, **15**, **21**, 22, 24, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, **45**
X_{inC}	**1**, **4**, **39**, **40**, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63

In order to analyze the influence of significance thresholds setting on the performance of the proposed method, the value of λ_{remove} is varied in the range of (0, 0.5]. The corresponding Ω_{p} and Ω_{r} under different significance thresholds are shown in Fig. 8.

When the significance threshold is raised, the conditions for adding the independent variables to the regression model will be more relaxed, so the recall rate will continue to increase, and the corresponding precision rate will continue to decrease in Fig. 8. The typical values are shown in Table IV below. According to $\Omega_{\text{p,ave}}|_{g=0.8}=95.1\%$, it can be observed that when more than 80% of the consumer phase connectivity can be identified, the average accuracy rate is higher than 95%.

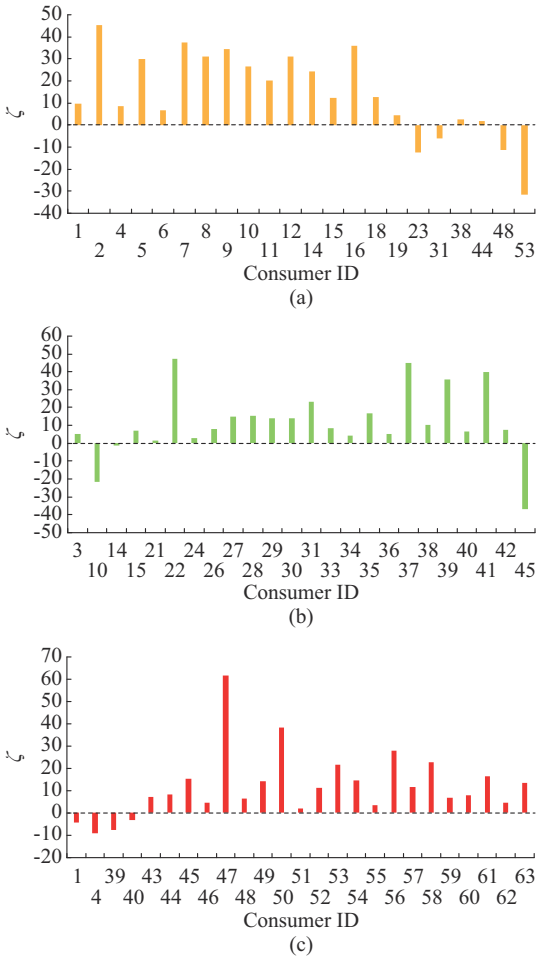

Fig. 6. Significance correction factors for variables of each phase. (a) Phase A. (b) Phase B. (c) Phase C.

To evaluate the computational burden, the SR algorithm implemented by MATLAB statistics toolbox (version R2019b) through the function "stepwisefit" is applied 100 times when λ_{remove} varies in the range of (0, 0.5]. The computational time of the SR process is counted and averaged.

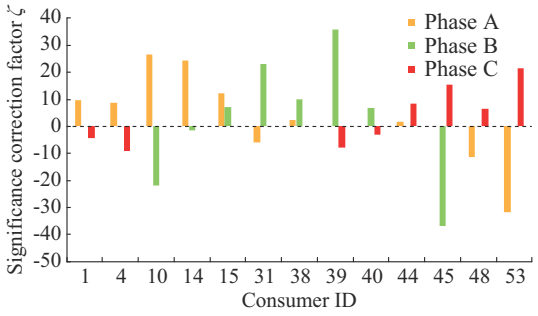

Fig. 7. Significance correction factors for variables in intersection.

TABLE III
CORRECTED IDENTIFICATION RESULTS USING ALGORITHM 2

Subset	Consumer ID
X'_{inA}	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 18, 19
X'_{inB}	3, 21, 22, 24, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42
X'_{inC}	43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63

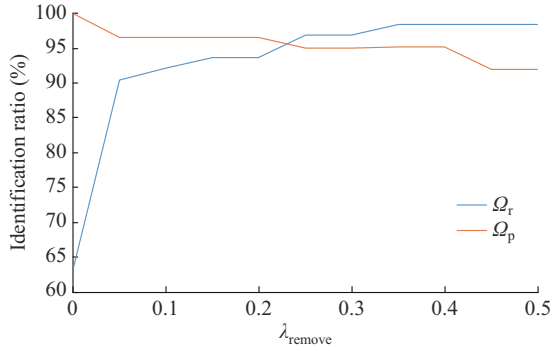

Fig. 8. Ω_{p} and Ω_{r} under different significance thresholds.

TABLE IV
Ω_{p} AND Ω_{r} UNDER DIFFERENT SIGNIFICANCE THRESHOLDS

λ_{remove}	λ_{entry}	Ω_{p} (%)	Ω_{r} (%)	$\Omega_{\text{p,ave}}\vert_{g=0.8}$ (%)
0.001	0.0005	100	63.5	95.1
0.500	0.2500	91.9	98.4	95.1

On average, the computation time used by the SR algorithm is approximately 32 s using a computer with an Intel Core i5-8265 CPU of 3.4 GHz and a RAM of 8 GB. Then, the computational time will increase insignificantly when the number of consumers in the network increases. Therefore, the proposed method is not applied to real-time applications.

C. Error Analysis

When the influence of hidden error is not considered, the relative error e/y of the regression model (5) is mainly affected by the model error and measurement error, wherein the model error is related to the type of the selected regression variable. Taking phase A of the LVDN shown in Fig. 4 as an example, under different measurement errors, the Monte Carlo method is used to randomly perform 5000 power flow simulations, and the cumulative distribution function curves based on the relative error e/y of current and active power are obtained, as shown in Fig. 9.

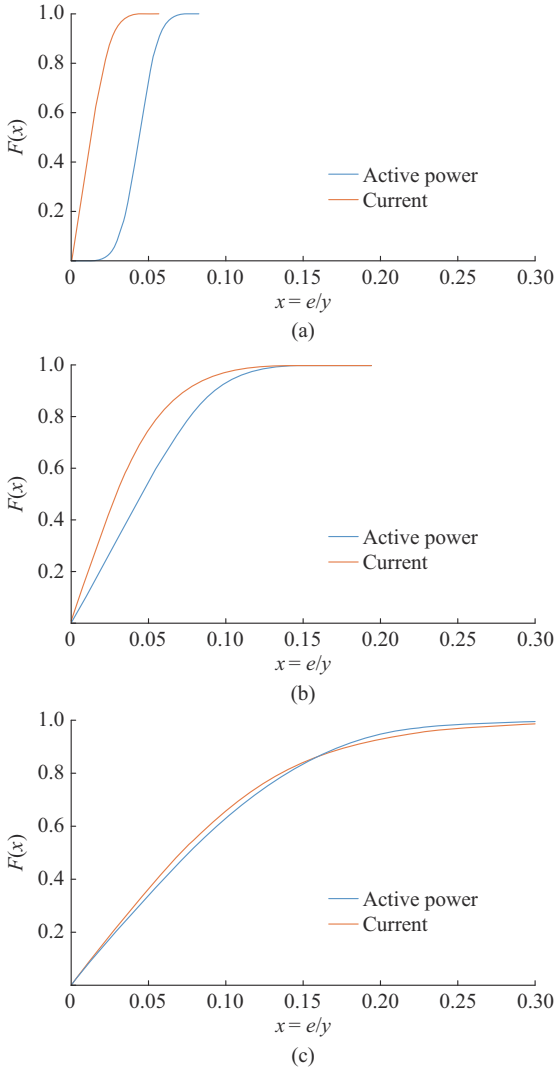

Fig. 9. Cumulative distribution function curves of e/y under different measurement errors. (a) $\varepsilon_s = 1\%$. (b) $\varepsilon_s = 4\%$. (c) $\varepsilon_s = 8\%$.

As can be observed from Fig. 9, when the measurement error ratio $\varepsilon_s = 1\%$, the relative error e/y generated based on the current and active power calculation does not exceed 7%; when the measurement error ratio $4\% \le \varepsilon_s \le 8\%$, the relative error has a large random variation range and can be up to 30%. From a statistical point of view, the relative error generated by the current-based calculation is likely to be smaller than that by the active power calculation. It means that within the normal measurement error range ($\varepsilon_s \le 8\%$), ignoring the phase angle produces less model error than ignoring the technical loss.

In order to further compare the identification results when the current and active power are used as regression variables, the test system shown in Fig. 4 is taken as an example. Under different measurement errors, the indices for $\Omega_{\text{p,ave}}\vert_{g=0.8}$ are shown in Table V. It can be observed that compared with the active power, using the current as the regres-

sion variable will be more beneficial in improving the identification accuracy of the proposed method.

TABLE V
$\Omega_{\text{p,ave}}|_{g=0.8}$ WITH DIFFERENT REGRESSION VARIABLES

ε_s (%)	$\Omega_{\text{p,ave}}\|_{g=0.8}$ (%)	
	Current	Active power
1	100.0	100.0
4	99.9	98.4
8	95.1	93.7

D. Comparison with Other Methods

This subsection compares the proposed method (M4) with LS method (M1) [31], the integer quadratic programming (IOP) method (M2) [18], Lasso regression method (M3) [22], and voltage correlation comparison (M5) [8]. Two types of scenarios are considered according to whether the hidden error is considered. The current is used as the optimization or regression variable and the identification results of each method are compared.

1) Without Considering Hidden Errors

Under different measurement errors, the precision rate Ω_{p} and recall rate Ω_{r} of M1, M2, M4, and M5 are calculated as shown in Table VI. Since M1, M2, and M5 have no independent variable screening mechanism, their recall rates are 100%.

TABLE VI
Ω_{p} AND Ω_{R} OF M1, M2, M4, AND M5 IN SCENARIO 1

ε_s (%)	M1		M2		M4		M5	
	Ω_{p} (%)	Ω_{r} (%)	Ω_{p} (%)	Ω_{r} (%)	Ω_{p} (%)	Ω_{r} (%)	Ω_{p} (%)	Ω_{r} (%)
1	98.41		100.0		83.2		100.0	100.0
4	96.80	100	98.4	100	77.5	100	100.0	93.7
8	85.70		95.2		70.2		96.5	90.5

Comparing the identification results of different methods, it can be observed that under the same measurement error, the proposed method has the highest accuracy. But when the measurement error is large ($\varepsilon_s \geq 4\%$), the recall rate of the proposed method cannot reach 100% due to the large measurement error. The accuracy corresponding to M5 is nonideal since the voltage profiles of customers within short electrical distances are similar.

2) Considering Hidden Error

One case is that the current observations of consumers 9-39 are either 0 or missing due to electricity theft or interruption of communication at partial time instants. Then, under different measurement errors, the precision rate of M1 and M2 are calculated, as shown in Table VII. The credible precision rates $\Omega_{\text{p,ave}}|_{g=0.8}$ for M3 and M4 are calculated as shown in Table VIII.

It can be observed from Table VII that when there is a hidden error and the measurement error is large ($\varepsilon_s > 4\%$), the precision rate of the traditional optimization methods will be lower than 70%, and can be as low as 63% in the extreme case. Such a low identification accuracy will have very limited practical applications.

TABLE VII
Ω_{p} OF M1 AND M2 IN SCENARIO 2

ε_s (%)	Ω_{p} (%)	
	M1	M2
1	80.9	73.0
4	79.3	69.8
8	66.7	63.5

TABLE VIII
$\Omega_{\text{p,ave}}|_{g=0.8}$ OF M3 AND M4 IN SCENARIO 2

ε_s (%)	$\Omega_{\text{p,ave}}\|_{g=0.8}$ (%)	
	M3	M4
1	82.7	98.2
4	81.7	95.2
8	71.5	84.4

It can be observed from Table VIII that when the test system has hidden errors and the measurement error is small ($\varepsilon_s \leq 4\%$), the proposed method can still ensure that the recall rate is not less than 80%, and the precision rate is greater than 95%. This is much larger than the precision rate of M3 under the same conditions ($\Omega_{\text{p}} = 81.7\%$). When the measurement error is large ($\varepsilon_s > 4\%$), the proposed method can still maintain a higher precision.

Another case is that the phase current observations measured by the sensors are missing at partial time instants. To avoid making mistakes in algorithm operation, these time instants with data missing should be abandoned. Considering the reduction of valid data samples, it could lead to the deterioration of results.

When the measurement error ratio is 4%, the precision rate Ω_{p} and recall rate Ω_{r} of M3 and M4 can be presented, as shown in Fig. 10. The range of (0, 0.5] is considered for the significance threshold λ_{remove} and the regularization parameter λ_{lasso} of M3.

It can be observed from Fig. 10 that when the algorithm parameters change, the precision rates of the proposed method and M3 are slightly changed but with a significant change of recall rates. When the recall rate is greater than 80%, the result accuracy of the proposed method is about 95%; while the accuracy of M3 is only about 80%. Thus, it can be observed that under the same recall rate, the accuracy of the proposed method is higher.

V. CONCLUSION

This paper has proposed an SR-based phase identification method of LVDN. The SR algorithm is used to identify the consumer phase connectivity based on their significances, and the significance correction factor is proposed for result correction. Through case studies based on a test system, the following conclusions can be drawn.

1) Compared with the LS and IQP methods, the proposed method has higher identification accuracy, especially when there is a hidden error. But the recall rate of the proposed method cannot reach 100% when the errors are large.

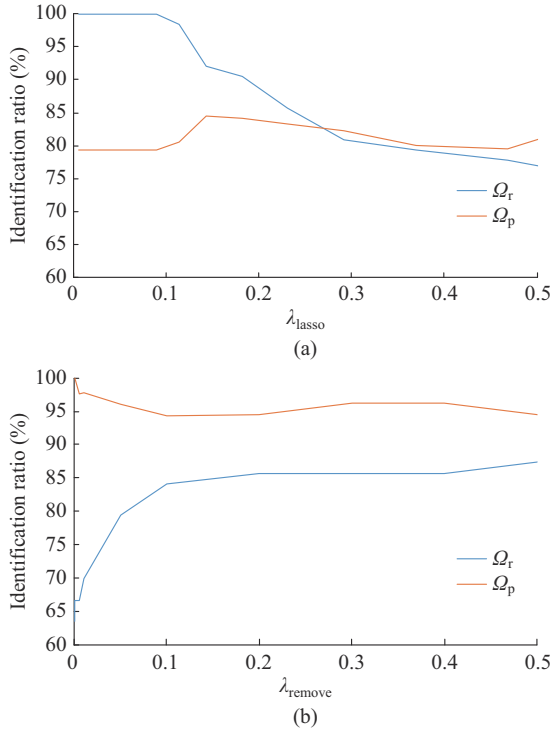

Fig. 10. Ω_p and Ω_r curves of M3 and M4. (a) M3. (b) M4.

2) Compared with using active power, when the current is used as the regression variable, the multi-linear model error is smaller, which is beneficial for improving the accuracy of the proposed method.

3) Compared with the Lasso regression method, when there is a hidden error and the recall rate is higher than 80%, the accuracy of the proposed method is increased by an average of 14%.

For the practical application of the proposed method, future research will focus on improving the recall rate without reducing the precision rate. Also, the impact of distributed generation on the identification results of the proposed method will also be studied.

REFERENCES

[1] F. Shen, Q. Wu, and Y. Xue, "Review of service restoration for distribution networks," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 1, pp. 1-14, Jan. 2020.

[2] R. Li, P. Wong, K. Wang *et al*., "Power quality enhancement and engineering application with high permeability distributed photovoltaic access to low-voltage distribution networks in Australia," *Protection and Control of Modern Power Systems*, vol. 5, no. 3, pp. 1-7, Aug. 2020.

[3] H. Yu, Y. Wu, W. Guan *et al*., "Practical method for data-driven user phase identification in low-voltage distribution networks," *Frontiers in Energy Research*, vol. 9, pp. 1-7, Nov. 2021.

[4] F. Therrien, L. Blakely, and M. J. Reno, "Assessment of measurement-based phase identification methods," *IEEE Open Access Journal of Power and Energy*, vol. 8, pp. 128-137, Mar. 2021.

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

[6] H. Pezeshki and P. J. Wolfs, "Consumer phase identification in a three phase unbalanced LV distribution network," in *Proceedings of 2012 3th IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, Berlin, Germany, Oct. 2012, pp. 1-7.

[7] H. Pezeshki and P. Wolfs, "Correlation based method for phase identification in a three phase LV distribution network," in *Proceedings of 2012 22th Australasian Universities Power Engineering Conference (AUPEC)*, Bali, Indonesia, Sept. 2012, pp. 1-7.

[8] W. Luan, P. Peng, M. Maras *et al*., "Smart meter data analytics for distribution network connectivity verification," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1964-1971, Jun. 2015.

[9] T. A. Short, "Advanced metering for phase identification, transformer identification, and secondary modeling," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 651-658, Jun. 2013.

[10] J. Peppanen, M. J. Reno, R. J. Broderick *et al*., "Distribution system model calibration with big data from AMI and PV inverters," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2497-2506, Sept. 2016.

[11] V. C. Cunha, W. Freitas, F. C. L. 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.

[12] B. Foggo and N. Yu, "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, Nov. 2018.

[13] B. Foggo and N. Yu, "Improving supervised phase identification through the theory of information losses," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2337-2346, May 2020.

[14] V. Arya and R. Mitra, "Voltage-based clustering to identify connectivity relationships in distribution networks," in *Proceedings of 2013 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Vancouver, Canada, Oct. 2013, pp. 7-12.

[15] S. Liu, X. Cai, Z. Lin *et al*., "Practical method for mitigating three-phase unbalance based on data-driven user phase identification," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1653-1656, Mar. 2020.

[16] L. Chao, Z. Lei, and Y. Li, "Topology checking method for low voltage distribution network based on fuzzy C-means clustering algorithm," in *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China, Jun. 2020, pp. 1077-1080.

[17] H. Di, Y. Xia, Q. Tao *et al*., "Dual learning for machine translation," in *Proceedings of 2016 30th Conference on Neural Information Processing Systems*, Barcelona, Spain, Nov. 2016, pp. 820-828.

[18] V. Arya, D. Seetharam, S. Kalyanaraman *et al*., "Phase identification in smart grids," in *Proceedings of 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Brussels, Belgium, Oct. 2011, pp. 25-30.

[19] P. Kumar and V. Arya, "Leveraging DERs to improve the inference of distribution network topology," in *Proceedings of 2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Dresden, Germany, Oct. 2017, pp. 52-57.

[20] M. Farajollahi, A. Shahsavari, and H. Mohsenian-Rad, "Topology identification in distribution systems using line current sensors: an MILP approach," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1159-1170, Mar. 2020.

[21] B. Appasani, A. V. Jha, 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. 124-135, Mar. 2021.

[22] X. Tang and J. V. Milanovic, "Phase identification of LV distribution network with smart meter data," in *Proceedings of 2018 IEEE PES General Meeting (PESGM)*, Portland, USA, Aug. 2018, pp. 1-5.

[23] 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.

[24] 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.

[25] S. J. Pappu, N. Bhatt, 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.

[26] M. Zhang, W. Luan, S. Guo *et al*., "Topology identification method of distribution network based on smart meter measurements," in *Proceedings of 2018 China International Conference on Electricity Distribution (CICED)*, Tianjin, China, Sept. 2018, pp. 372-376.

[27] MathWorks. (2022, Oct.). MATLAB Statistics Toolbox User's Manual, The MathWorks [Online]. Available: https://ww2. mathworks. cn/products/statistics.html

[28] N. Zhou, J. W. Pierre, and D. Trudnowski, "A stepwise regression method for estimating dominant electromechanical modes," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 1051-1059, May 2012.

[29] L. Shao, Y. Hu, and G. Xu, "A high precision on-line detection method for IGBT junction temperature based on stepwise regression algorithm," *IEEE Access*, vol. 8, pp. 186172-186180, Jan. 2020.
[30] F. Wei, Y. Cai, and J. Tang, "Low-voltage station area topology recognition method based on weighted least squares method," in *Proceedings of 2020 IEEE Sustainable Power and Energy Conference (iSPEC)*, Chengdu, China, Nov. 2020, pp. 2539-2544.
[31] L. Zhou and Y. Zhang, "Consumer phase identification in low-voltage distribution network considering vacant users," *International Journal of Electrical Power & Energy Systems*, vol. 121, pp. 1-11, Apr. 2020.
[32] L. Zhou and Q. Li, "Consumer phase identification under incomplete data condition with dimensional calibration," *International Journal of Electrical Power & Energy Systems*, vol. 129, pp. 1-13, Feb. 2021.
[33] *Electromagnetic Compatibility (EMC) Part 4-30: Testing and Measurement Techniques – Power Quality Measurement Methods*, IEC 61000-4-30 ED. 3.0, 2021.
[34] C. Si, S. Xu, 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.
[35] 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.
[36] 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.

Yingqi Yi received the M.Sc. 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 data analytics in distribution system and optimal operation of distribution network.

Siliang Liu received the M.Sc. 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.

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.

Ying Xue received the B.Eng. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, and the University of Birmingham, Birmingham, U.K., in 2012, and the Ph.D. degree in electrical engineering from the University of Birmingham, in 2016. His main research interests include modelling, control and simulation of HVDC, and renewable generation systems.

Wenyang Deng received the M.Sc. 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. He is now a Postdoctoral Researcher at the South China University of Technology, Guangzhou, China. His main research interests include regulation of inverters, power sharing and power quality improvement in microgrids.

Qinhao Li received the B.E. and Ph.D. degrees in electrical engineering from the South China University of Technology, Guangzhou, China, in 2012 and 2018, respectively. He is currently a Postdoctoral Researcher with the School of Electric Power, South China University of Technology, Guangzhou, China. His main research interests include reactive power optimization, renewable energy integration, and power Internet of Things.