Fault Location in Low Voltage Smart Grids Based on Similarity Criteria in the Principal Component Subspace

Laiz Souto
Control Engineering and Intelligent Systems
Universitat de Girona
Girona, Girona, Spain
laiz.souto@gmail.com

Joaquim Meléndez, Sergio Herraiz
Control Engineering and Intelligent Systems
Universitat de Girona
Girona, Girona, Spain
joaquim.melendez,sergio.herraiz@udg.edu

Abstract—**This paper presents a new strategy based on multivariate statistical analysis for fault location and classification in power distribution networks with distributed energy resources, variable loads, and switches enabling grid reconfiguration. The statistical method relies on impedance measurements acquired at the substation buses to build a data-driven model of the network operating conditions with dimensionality reduction, and considers a few reference scenarios representing standard operating conditions and short-circuit operation to perform fault location and classification with use of similarity criteria in the principal component subspace. Moreover, this paper includes a case study with a real-based low voltage power distribution network to test and validate the methodology.**

Index Terms—**fault location, machine learning algorithms, power distribution faults, smart grids, statistical learning**

I. Introduction

The increasing application of digital technology is gradually transforming old-fashioned power distribution networks into modern smart grids with enhanced supervision, protection, and control features. Nevertheless, fault location and classification tasks still face challenges related to limited measurements along the feeders, usually available only at the distribution substation, and the inaccurate representation of network components, such as loads, distributed generators, lines, and the status of switches [1]. As a result, many faults are identified correctly only after trouble calls from affected customers, which may lead to unacceptable interruption times and have negative impacts for the system operation in general. However, in the event of a fault, information about its type and location should be available as soon as possible to start grid reconfiguration and restore normal energy supply [2]. Therefore, an automated strategy capable of overcoming these stumbling

blocks is necessary for fast, accurate, adequate fault location and classification.

In this scenario, there are great opportunities for artificial intelligence (AI) techniques, as the increased processing power and reduced costs of computers enable the application of cutting-edge mathematical and information processing strategies in the search of faults. A few AI-based approaches for fault location at distribution level have been proposed recently; for instance, [3] combines principal component analysis, support vector classifiers, and feed-forward neural networks to perform fault location and classification in radial distribution networks, using measurements available at the substation together with information about circuit breaker and relay statuses; [4] presents a data-driven mixed-integer linear programming algorithm for fault location relying on smart meters at low voltage (LV) level and remote fault indicators at medium voltage (MV) level; [5] presents a feature selection method based on the information gain and minimum description length discretization algorithm together with a complementary expert information system to detect high-impedance faults; [6] uses continuous wavelet transform to generate gray-scale images of transient zero-sequence current signals together with a convolutional neural network for feature extraction and fault detection in resonant grounding distribution systems; [7] applies the Stockwell transform to three-phase current signals and extracts features used as inputs in different machine learning tools with the goal of locating different types of faults in power distribution grids; [8] reduces the multiple estimation problem in fault location by combining support vector machines and the k-nearest neighbors with features extracted from fundamental voltage and current signals; and [9] applies the Fischer-Rao registration method to preserve the shape of data from faults at different locations and operating conditions and hierarchical cluster analysis for fault classification.

Fitting into this context, this paper presents a new strategy for fault location and classification in power distribution networks with distributed energy resources and variable loads installed along the feeders that is general enough to consider

changes in the grid configuration (e.g. switch status). Thus, it assumes that phasor measurement units (PMU) or similar measurement devices are deployed at the secondary of substation transformers. The methodology relies on impedance measurements gathered at the substation buses to build a statistical model of the network operating conditions with dimensionality reduction and considers a few reference scenarios representing standard operation and short-circuits to perform fault location and classification based on similarity criteria in the principal component subspace. Moreover, it is capable of distinguishing faults from variations in the standard operating conditions and identifying the grid configuration correctly. Furthermore, testing is conducted in a real-based LV power distribution network under different fault conditions.

This text is structured as follows: Section II presents the methodology, Section III describes the application example used to test and validate the method, Section IV includes simulation results and discussions, and Section V presents the conclusions.

II. METHODOLOGY

This section introduces the multivariate statistical analysis used to build a data-driven model of the network operating conditions with dimensionality reduction and provides the theoretical background necessary to go through the rest of this paper. For an in-depth explanation, see [10].

First, consider a set K of reference scenarios in which distinct operating conditions of the power distribution network under consideration are represented (i.e. including standard operation and short-circuits). Then, let $\mathbf{X}_{(k)}$ be the $n \times m$ observation matrix (1) of the kth scenario, centered (zero mean) and scaled (unit variance), with n observations referred to the number of phasor quantities sampled over time and m variables referred to measurements of phasor quantities at every substation.

$$\mathbf{X}_{(k)} = \begin{bmatrix} x_{1,1(k)} & x_{1,2(k)} & \cdots & x_{1,m(k)} \\ x_{2,1(k)} & x_{2,2(k)} & \cdots & x_{2,m(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1(k)} & x_{n,2(k)} & \cdots & x_{n,m(k)} \end{bmatrix} \quad (1)$$

The covariance matrix $\mathbf{S}_{(k)}$ can be computed from $\mathbf{X}_{(k)}$ and further decomposed in the $m \times m$ matrices $\mathbf{V}_{(k)}$ and $\mathbf{\Lambda}_{(k)}$ using eigenvalue decomposition according to (2). Columns in $\mathbf{V}_{(k)}$ are the eigenvectors and contain the principal components, which represent orthonormal vectors whose directions express the major variability of the data and the relative weights of the original variables. In turn, $\mathbf{\Lambda}_{(k)}$ is a diagonal matrix and contains the eigenvalues, which express variability in the direction of each principal component or column of $\mathbf{V}_{(k)}$. The matrices $\mathbf{V}_{(k)}$ and $\mathbf{\Lambda}_{(k)}$ can be written as (3) and (5), respectively.

$$\mathbf{S}_{(k)} = \frac{1}{n-1}\mathbf{X^T}_{(k)}\mathbf{X}_{(k)} = \mathbf{V}_{(k)}\mathbf{\Lambda}_{(k)}\mathbf{V^T}_{(k)} \quad (2)$$

$$\mathbf{V}_{(k)} = \begin{bmatrix} \mathbf{v_{1(k)}} & \mathbf{v_{2(k)}} & \cdots & \mathbf{v_{m(k)}} \end{bmatrix} \quad (3)$$

with

$$\mathbf{v}_{.(\mathbf{k})} = \begin{bmatrix} v_{1,.(k)} & v_{2,.(k)} & \cdots & v_{m,.(k)} \end{bmatrix}^T \quad (4)$$

$$\mathbf{\Lambda}_{(k)} = \begin{bmatrix} \lambda_{1(k)} & 0 & \cdots & 0 \\ 0 & \lambda_{2(k)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{m(k)} \end{bmatrix} \quad (5)$$

Once $\mathbf{V}_{(k)}$ and $\mathbf{\Lambda}_{(k)}$ are computed, dimensionality reduction is achieved by retaining $r < m$ principal components or columns of $\mathbf{V}_{(k)}$ which present the largest eigenvalues $\lambda_{1(k)}, \cdots, \lambda_{r(k)}$. As a result, $\mathbf{V}_{(k)}$ is reduced to an $m \times r$ matrix $\mathbf{P}_{(k)}$ given by (6) which represents the major trends of the data set with some loss of information.

$$\mathbf{P}_{(k)} = \begin{bmatrix} \mathbf{v_{1(k)}} & \mathbf{v_{2(k)}} & \cdots & \mathbf{v_{r(k)}} \end{bmatrix} \quad (6)$$

Next, consider a generic testing scenario denoted by k', possibly containing a fault or some deviation from the standard operating conditions, and let $\mathbf{V}_{(k')}$ and $\mathbf{\Lambda}_{(k')}$ be its eigenvector matrix and eigenvalue matrix, respectively. For each k' under consideration, the choice of an appropriate value of r to reduce $\mathbf{V}_{(k')}$ into $\mathbf{P}_{(k')}$ is based on the similarity criteria, calculated as a weighted cosine sum $\varphi_{k,k'}$ of the dot product of $\mathbf{v_{j(k)}}$ and $\mathbf{v_{j(k')}}$ weighted by the normalized variance $\bar{\lambda}_{j,j(k')}$, $j = 1, \cdots, r$, as in (7). Only the the reference scenarios $k \in K_0$ with standard operation or faults at the secondary substation buses are considered to select r principal components. This procedure also allows to identify the correct network configuration and operating condition of the testing scenario, since $\varphi_{k,k'}$ should be close to 1.0 if $\mathbf{v_{j(k)}}$ and $\mathbf{v_{j(k')}}$, $j = 1, \cdots, r$, are similar.

$$\varphi_{k,k'} = \sum_{j=1}^{r} |\mathbf{v_{j(k)}} \cdot \mathbf{v_{j(k')}}| \, \bar{\lambda}_{j,j(k')} \quad (7)$$

with

$$\bar{\lambda}_{j,j(k')} = \frac{\lambda_{j,j(k')}}{\sum_{j=1}^{m} \lambda_{j,j(k')}} \quad (8)$$

Once r is defined, if a fault occurred, further investigation is conducted to identify its possible locations. In this case, fault location and classification are performed by comparing the statistical model given by $\mathbf{P}_{(k')}$ and $\mathbf{\Lambda}_{(k')}$ with those computed for the $k \in K - K_0$ reference scenarios not evaluated previously. The results calculated with (7) are ranked in descending order. Finally, the fault buses of the training scenarios with the highest values of (7) are identified as the most probable locations of the fault in the testing scenario.

III. CASE STUDY

The methodology was tested in a real-based LV power distribution network simulated in Matrix Laboratory (MATLAB) illustrated in Fig. 1. It represents a LV distribution network located in Catalonia, Spain, which consist of primary distribution
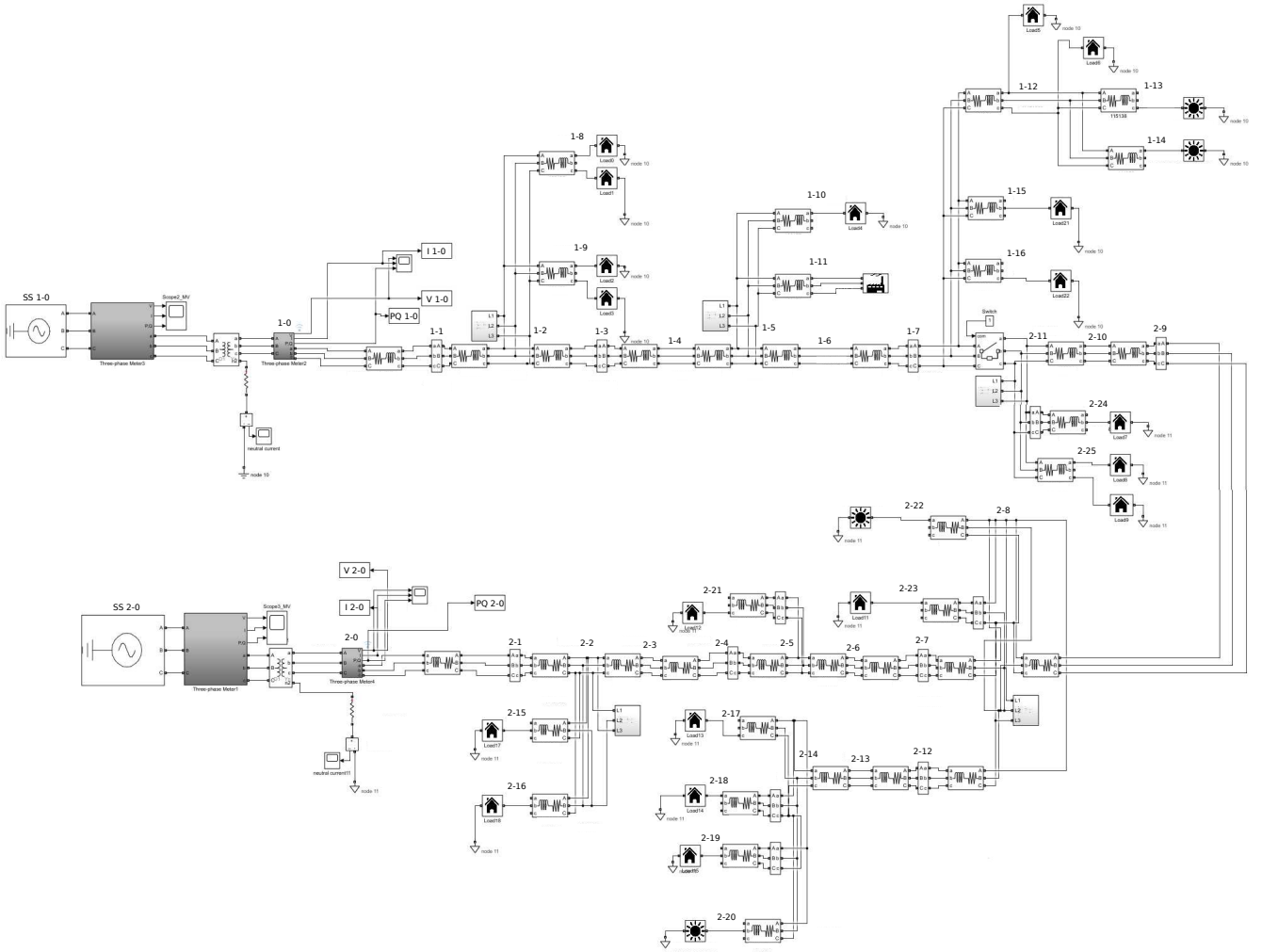
Fig. 1. Diagram of the low voltage distribution network modeled in Simulink, with 2 distribution substations (SS-1 − 0 and SS-2 − 0), 41 feeder sections, 1 switch, 4 PV panels, 1 industrial three-phase customer, and 19 residential single-phase customers.

feeders with branches connecting the substation node to the customers (i.e. local energy producers or consumers). In total, the network has 43 buses, 2 distribution substations (one with a 250-kVA and the other with a 630-kVA transformer Dyn11, 400 V secondary), 41 feeders modeled as short R-L lines (with $\frac{R}{X} = 5.4$ for overhead lines and $\frac{R}{X} = 2.7$ for underground cables), 1 switch, 20 different energy consumers (among them, 1 industrial, three-phase customer with 70 kW of contracted power and and 19 residential, single-phase customers with less than 10 kW of contracted power), and distributed generation from 4 solar photovoltaic (PV) modules (10 kWp each). In total, the length of the primary distribution feeder connecting SS-1 − 0 to SS-2 − 0 is 325 m and the length of the longest lateral branch is 95 m.

PMUs are installed at both substation nodes and sample phase voltage and line current phasor quantities from which the equivalent impedance is calculated. In other words, the number of variables is $m = 2 \times 3 = 6$ in all scenarios. It is noteworthy that the statistical models are built with impedance magnitudes only to suit the algorithm in use, as it is linear.

This choice is not expected to make a negative impact on the accuracy of fault location, since the network behavior and loads are mainly resistive and the faults are purely resistive.

In the reference scenarios, standard operation and faults are simulated with typical hourly values of PV generation and load consumption profiles over a year, which provides $n = 365$ observations per reference scenario. Although this time intervals are chosen due to the real PV generation and load consumption profiles available for the simulations, it is noteworthy that a shorter or longer time interval can be used to build the statistical models without loss of generality. In addition, the reference fault scenarios include three-phase symmetrical faults with fault resistance $R_F = 1m\Omega$ applied at the substation and load buses at midday. Both switch statuses on and off are considered in all training and testing scenarios.

In turn, testing scenarios consist of variations in the PV generation under normal operation, reduced by 25%, 50%, 75%, and 100% of the standard operation profiles, and three-phase symmetrical faults with fault resistance $R_F = 1\Omega$ simulated at the load buses at midday, considering the same
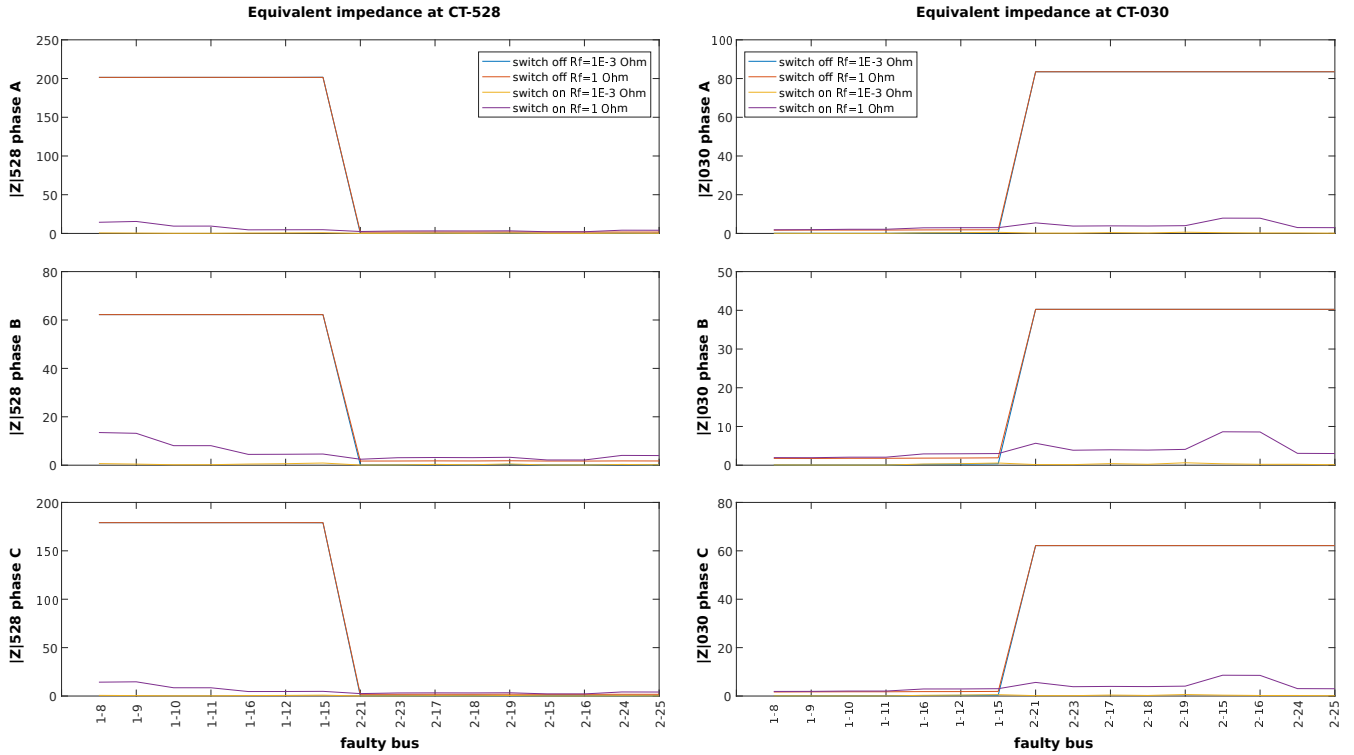
Fig. 2. Equivalent short-circuit impedance seen at SS-1 − 0 and SS-2 − 0 with three-phase faults at different buses.

PV generation and load consumption profiles as the reference scenarios. In total, 32 fault scenarios are tested, since buses 1 − 8, 1 − 9, 1 − 12, and 2 − 25 contain two households each (i.e. 16 faulty buses times 2 switch modes), besides 4 scenarios with variations in PV generation profiles under normal operation. For illustration and comparison purposes, the equivalent short-circuit impedance seen at the secondary substation buses 1 − 0 and 2 − 0 in the training and testing scenarios with faults ($R_F = 1m\Omega$ and $R_F = 1\Omega$, respectively) is plotted in Fig. 2.

IV. RESULTS AND DISCUSSION

The deviations computed under normal operation with variations in the PV generation profiles are displayed in Table I, considering standard operation and faults at the substation buses. Meanwhile, the fault location results obtained with the switch off and on are displayed in Table II and Table III, respectively, which describe the fault scenario k' (i.e. faulty bus) in the first column and the results obtained with the similarity criteria in the principal component subspace, including the correct identification of the grid setting regarding the switch mode and part of the network where the fault is in the second column; the faulty bus $k \in K$ determined by considering the sum of dot products calculated with (7) for all candidate buses, given information about the right network setting, in the third column; the distance error between the actual fault bus k' and the calculated bus k in the fourth column; and the r principal components used to compute (7) for all candidate scenarios in the fifth column.

TABLE I
CALCULATED RESULTS OF (7) WITH VARIATIONS IN THE PV GENERATION

| Testing scenario | Reference scenario | | | | | |
| | Std. operation | | Fault 1 − 0 | | Fault 2 − 0 | |
Switch mode	off	on	off	on	off	on
75% PV switch off	0.8254	0.6126	0.0350	0.5056	0.0381	0.1426
75% PV switch on	0.7567	0.9996	0.6860	0.8864	0.1442	0.3411
50% PV switch off	0.8244	0.6355	0.0725	0.5271	0.0264	0.1179
50% PV switch on	0.7557	0.9995	0.6865	0.8856	0.1408	0.3435
25% PV switch off	0.7954	0.6385	0.1065	0.5340	0.0146	0.0865
25% PV switch on	0.7558	0.9995	0.6863	0.8852	0.1398	0.3435
0% PV switch off	0.7944	0.6587	0.1472	0.5491	0.0040	0.0651
0% PV switch on	0.7512	0.9991	0.6908	0.8834	0.1340	0.3505

TABLE II
FAULT LOCATION RESULTS WITH SWITCH MODE OFF

Faulty bus (k')	Grid setting (Right/wrong)	Distance errors Bus (k)	k-k' [m]	r
1-8	Right	1-8	0.0	2
1-9	Right	1-8	20.6	2
1-10	Right	1-11	3.8	4
1-11	Right	1-8	58.5	2
1-12	Right	1-11	98.1	2
1-15	Right	1-11	121.0	2
1-16	Right	1-8	140.7	4
2-15	Right	2-17	194.2	3
2-16	Right	2-24	122.1	2
2-17	Right	2-24	94.2	2
2-18	Right	2-25	109.2	4
2-19	Right	2-24	117.8	2
2-21	Right	2-17	146.9	4
2-22	Right	2-17	90.9	4
2-24	Right	2-24	0.0	2
2-25	Right	2-17	89.6	2

TABLE III
FAULT LOCATION RESULTS WITH SWITCH MODE ON

Faulty bus (k')	Grid setting (Right/wrong)	Distance errors		r
		Bus (k)	k-k' [m]	
1-8	Right	1-8	0.0	2
1-9	Right	1-9	0.0	2
1-10	Right	1-9	57.5	4
1-11	Right	1-9	57.5	2
1-12	Right	1-9	151.8	3
1-15	Right	1-9	174.7	2
1-16	Right	1-9	139.7	4
2-15	Right	2-22	107.2	2
2-16	Right	2-22	101.2	2
2-17	Right	2-25	109.2	2
2-18	Right	2-18	0.0	4
2-19	Right	2-25	132.8	2
2-21	Right	2-22	60.5	4
2-22	Right	2-22	0.0	4
2-24	Right	2-22	24.7	3
2-25	Right	2-21	79.8	2

The results displayed in Tables I to III indicate that the multivariate statistical case-based reasoning strategy is capable of distinguishing between faults, standard operation, and variations in the standard operating conditions correctly. Additionally, in the event of a fault, the methodology identifies the part of the network where it occurred correctly in all cases and the true location of the fault with good accuracy in most cases. The fault location is identified correctly in 2 out of 16 scenarios when the switch is off and in 4 out of 16 scenarios when the switch is on, whereas the maximum distance error is $194.2\,m$ when the switch is off (faulty bus $2-15$ identified as $2-17$) and $174.7\,m$ when the switch is on (faulty bus $1-15$ identified as $1-9$). Despite the correct identification of the network setting, these errors are approximately the double of the length of the longest lateral branch and respectively stand for 59.8% and 53.8% of the total length from one substation to the other. Nevertheless, the actual faulty bus is among the first ranked results in almost all scenarios with both switch modes on and off when only the right network setting is considered, which shows the importance of identifying the grid setting correctly before performing fault location with this methodology. Consequently, the maximum errors of this fault location procedure remain in the same part of the network delimited by the switch where the point of fault is. Moreover, the average errors of all scenarios listed in Tables II and III stand for 27.0% and 23.0% of the total length from one substation to the other, which is less than the length of the longest lateral branch. The overall results are acceptable, as the fault location problem is a typical multiple-solution problem. Furthermore, Fig. 2 evinces that the equivalent impedance seen at the substations is almost the same for faults at different buses in the same part of the network and a fixed grid configuration.

It is noteworthy that the ranking of results according to (7) may be inaccurate over a range of scenarios due to the differences between the training and testing scenarios in use, such as variations in the standard operating conditions, differ-

ent faults, inaccurate network representations, etc. Therefore, the method can be improved by including more reference scenarios in the training data sets, including different faults, timescales, operating conditions, and additional information about the network topology, data from different sources, etc.

V. CONCLUSIONS

The multivariate statistical case-based reasoning strategy presented in this article is capable of locating and classifying faults with good accuracy. Moreover, it is also capable of identifying the network configuration correctly and distinguishing faults from variations in the standard operating conditions. The procedure used to identify the correct grid configuration prior to the location of the fault improves the accuracy of the method, as it reduces the number of candidate scenarios and limits the search to the right part of the network in all testing scenarios. Nonetheless, the method may provide inaccurate results over a range of scenarios, since the fault location problem presents multiple solutions.

REFERENCES

[1] H. Mohsenian-Rad, E. Stewart, and E. Cortez, "Distribution synchrophasors: Pairing big data with analytics to create actionable information," *IEEE Power and Energy Magazine*, vol. 16, no. 3, pp. 26–34, May 2018.

[2] A. Zidan, M. Khairalla, A. M. Abdrabou, T. Khalifa, K. Shaban, A. Abdrabou, R. El-Shatshat, and A. M. Gaouda, "Fault detection, isolation, and service restoration in distribution systems: State-of-the-art and future trends," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2170–2185, Sep. 2017.

[3] D. Thukaram, H. P. Khincha, and H. P. Vijaynarasimha, "Artificial neural network and support vector machine approach for locating faults in radial distribution systems," *IEEE Transactions on Power Delivery*, vol. 20, no. 2, pp. 710–721, April 2005.

[4] Y. Jiang, "Data driven fault location of electric power distribution systems with distributed generation," *IEEE Transactions on Smart Grid*, pp. 1–1, 2019.

[5] Q. Cui, K. El-Arroudi, and Y. Weng, "A feature selection method for high impedance fault detection," *IEEE Transactions on Power Delivery*, vol. 34, no. 3, pp. 1203–1215, June 2019.

[6] M. Guo, X. Zeng, D. Chen, and N. Yang, "Deep-learning-based earth fault detection using continuous wavelet transform and convolutional neural network in resonant grounding distribution systems," *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1291–1300, Feb 2018.

[7] M. Shafiullah, M. A. Abido, and T. Abdel-Fattah, "Distribution grids fault location employing st based optimized machine learning approach," *Energies*, vol. 11, no. 9, 2018.

[8] J. Mora-Florez, G. Morales-Espana, and S. Perez-Londono, "Learning-based strategy for reducing the multiple estimation problem of fault zone location in radial power systems," *IET Generation, Transmission Distribution*, vol. 3, no. 4, pp. 346–356, April 2009.

[9] J. Cordova, R. Arghandeh, Y. Zhou, S. Wesolowski, W. Wu, and S. Matthias, "Shape-based data analysis for event classification in power systems," in *2017 IEEE Manchester PowerTech*, June 2017, pp. 1–6.

[10] E. Russell, L. Chiang, and R. Braatz, *Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes*, ser. Advances in Industrial Control. Springer London, 2000.