Enhancing Accuracy While Reducing Computation Complexity for Voltage-Sag-Based Distribution Fault Location

Yimai Dong, *Student Member, IEEE*, Ce Zheng, *Student Member, IEEE*, and Mladen Kezunovic, *Fellow, IEEE*

***Abstract*—A fault-location method for radial distribution systems is proposed in this paper. The proposed method uses voltage and current phasors from feeder root and voltage sags measured at sparse nodes along the feeder, and pinpoints faults to the nearest node. Decision-tree (DT)-based fault segment identification is introduced before the process of node selection to reduce the computational complexity and improve fault-location accuracy. The method has been implemented on a practical distribution system and tested under a large number of fault scenarios. Test results are compared with those from the traditional voltage-sag-based fault-location algorithm using the same inputs, and the conclusion is that the proposed method can achieve more reliable results while maintaining computational simplicity. A quantitative method to suggest the optimal placement of measurement units based on the DT variable importance is proposed at the end.**

***Index Terms*—Decision trees (DTs), fault location, optimal sensor placement, power distribution, voltage sags.**

I. Introduction

THE ACCURACY and computational complexity are the two most important criteria when evaluating a fault-location algorithm. The accuracy of fault-location results has a great impact on fault isolation and repair activities and, thus, the overall duration of fault-caused outage; the implementation of an algorithm may be restrained by its computational complexity [1]. Achieving accuracy while maintaining computational simplicity is challenging for distribution system-level fault location, because of the number of components, heterogeneity of lines, unbalanced operation, time-varying load condition, and most of all, lack of measurements [2].

Currently, there are two categories of fault-location techniques: outage mapping and precise location. Outage mapping is a group of techniques that intend to narrow down the area where the fault occurs, based on information from customer calls, circuit breaker (CB) status, advanced metering, and the geographic information system (GIS) model [3], [4]. Another category comprises techniques that determine the precise location of the fault through calculation using field measurements. Subcategories of precise location methods are impedance-based methods using sequential network analysis or direct circuit analysis [5]–[10]; frequency component-based methods [11]–[13]; and methods based on sparse voltage measurements and postfault power-flow analysis [14]–[16].

The most distinctive feature of voltage measurement-based methods is the capability of differentiating faults on different laterals with the same equivalent fault impedance seen from the beginning of a feeder. Despite the advantage, a major concern of such methods is their computational burden. The methods determine the location of the fault by assuming a fault on every tentative node, solving postfault power flow and comparing the calculated voltage sags with measured ones. Without an effective screening mechanism, the pool of tentative nodes usually contains all nodes on a feeder. Power flow is calculated by iterative procedures. The computational burden is in proportion to the multiplication of the number of tentative nodes and number of iterations. On the other hand, not every node in the system is observable due to the limited number of measurements, so the outputs of these methods are under the risk of large errors when two or more similar (in the sense of electric quantities) laterals exist in one unobservable area.

To deal with the lack of measurements, knowledge-based approaches are introduced to the field of fault processing. Among others, the decision-tree (DT) method was first introduced to the field of fault analysis in the 1990s. In [17], the DT is applied to the problem of fault diagnosis, in particular, the fault-type classification. In [18], Sheng *et al.* used DT to distinguish the high impedance fault from normal system operations. A review of literature reveals that although the DT was applied in several works to estimate the fault section [19], [20], the important issue of how DT can enhance the accuracy of existing fault-location algorithms has not yet been fully studied.

In this paper, a two-step fault-location algorithm is proposed. In step 1, a DT-based approach is introduced to determine the faulted segment; in step 2, an improved fault-location algorithm based on [15] is adopted to assess the likelihood of nodes belonging to the segment from step 1. The classification tree proposed by Breiman *et al.* [21] will be employed for fast fault segment estimation, and the performance of the fault-location algorithm aided by the DT method will be examined.

This paper is organized as follows: limitations of the traditional voltage measurement-based fault-location algorithm are discussed first in Section II. The formulation of the proposed

Manuscript received October 16, 2012; revised January 21, 2013; accepted February 07, 2013. Date of publication March 07, 2013; date of current version March 21, 2013. Paper no. TPWRD-01119-2012.

The authors are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128 USA (e-mail: dongyimai@tamu.edu; zhengce@tamu.edu; kezunov@ece.tamu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRD.2013.2247639

method is in Section III, including the knowledge-based segment selection and revised fault-location algorithm. Implementation procedures are detailed in Section IV, and case studies are given in Section V. In the end, a quantitative approach is proposed to suggest the optimal sensor placement for better fault-location estimation based on the DT variable importance.

II. Theoretical Background

A. Voltage Measurement-Based Fault-Location Methods

The voltage measurement-based method is first proposed by Galijasevic and Abur in [14], where the concept of vulnerability contours is used in assessing the likelihood of voltage sags affecting a given network area. In [15], Pereira *et al.* extended the formulation in [14] assuming the availability of voltage and current phasors at the feeder root, and voltage sag measurements from sensors along the feeder. Voltage sags were calculated using a postfault load-flow approach that does not require the estimation of fault resistance. In [16], Lotfifard *et al.* assumed postfault phase-angle shifts that were available from sparse measurements, and proposed an approach for eliminating some tentative nodes by characterizing the voltage sags from different sensors. A new index was proposed for analyzing voltage sags and angle shifts calculated from the load-flow computation based on estimated fault resistance.

B. Pereira's Algorithm [15]

The fault-location method from [15] is based on the fact that different drops in voltage amplitudes (voltage sags) are experienced by each feeder node during a fault. The algorithm runs the prefault load flow first, then assigns one node as the faulted node, runs postfault load flow, calculates voltage sags, and calculates the difference (mismatch) between calculated and measured values at measurement points in the system. When faults on all tentative nodes have been simulated, the tentative node with the smallest mismatch is selected as output.

The core of Pereira's algorithm is the calculation of load flows. An iterative load-flow algorithm for the radial distribution system described in [22] is used to solve prefault load flow. Back-sweeping to update branch currents using (1) and (2) and forward-sweeping to update node voltages using (3) are conducted in each iteration. The stopping criterion for iterations is defined

$$I_{j\_n^{(k)}} = {Z_{L\_n}}^{-1} \cdot V_n^{(k-1)} \tag{1}$$

$$I_{b\_i}^{(k)} = \sum I_{b\_p}^{(k)} + I_{j\_n}^{(k)} \tag{2}$$

$$V_n^{(k)} = V_m^{(k)} - Z_{b\_i} \cdot I_{b\_i}^{(k)} \tag{3}$$

$$\max\left\{\left|V_n^{(k)} - V_n^{(k-1)}\right|\right\} < \varepsilon,\ n = 1, \ldots, N \tag{4}$$

where

- k number of iterations;
- $I_{j\_n}^{(k)}$ injection current at node n;
- Z_{L_n} three-phase load impedance matrix at node n;

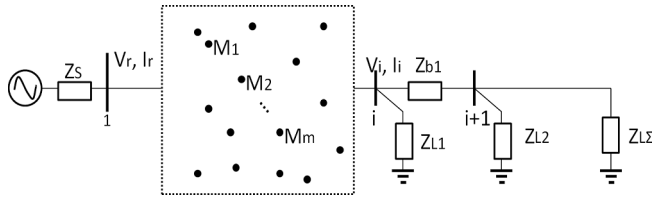


Fig. 1. One-line diagram of a feeder.

- $V_n^{(k)}$ node voltage of the downstream node of branch i;
- $I_{b\_i}^{(k)}$ branch current of branch i, which flows from node m to node n;
- $I_{b\_p}^{(k)}$ branch current of branch p, which flows out from node n;
- Z_{b_i} three-phase line impedance matrix for branch i;
- ε threshold for a change in node voltage;
- N total number of nodes.

In postfault load-flow computation, similar procedures are used, except that the mismatch between measured and calculated values of feeder current is calculated after calculation of branch currents (5), injected to the assumed faulted node (6), and the branch current is updated again using (2)

$$I_f^{(k)} = I_f^{(k-1)} + (I_r^{\mathrm{pf,meas}} - I_r^{\mathrm{pf,cal}}) \tag{5}$$

$$I_{j\_n}^{pf,(k)} = I_{j\_n,L}^{pf,(k-1)} + I_f^{(k)} \tag{6}$$

where

- $I_f^{(k)}$ fault current;
- $I_r^{\mathrm{pf,meas}}$ current measured at the feeder root;
- $I_r^{\mathrm{pf,cal}}$ calculated current at the feeder root;
- $I_{j\_n}^{\mathrm{pf,(k)}}$ injection current at faulted node n;
- $I_{j\_n,L}^{\mathrm{pf,(k-1)}}$ injection current from the load connected to n.

C. Limitations in Pereira's Algorithm

Pereira's approach smartly bypassed the estimation of fault resistance. However, it introduced confusion when no measurements were taken from the downstream of the faulted node. This can be explained by circuit analysis. Fig. 1 depicts such a case. V_r and I_r are voltage and current phasors at the feeder root. The dotted box represents the unfaulted part of the feeder, which contains all of the measurement nodes (M_1 to M_m). Z_{b1} is the branch impedance between node i and $i+1$. Z_{L1} and Z_{L2} are load impedance connected to node i and $i+1$. $Z_{L\Sigma}$ is the equivalent impedance of branches and loads behind node $i+1$.

The network between node 1 and node i can be represented as a two-port network

$$\begin{bmatrix} V_r \\ V_i \end{bmatrix} = \begin{bmatrix} Z_{rr} & Z_{ri} \\ Z_{ir} & Z_{ii} \end{bmatrix} \begin{bmatrix} I_r \\ I_i \end{bmatrix}. \tag{7}$$

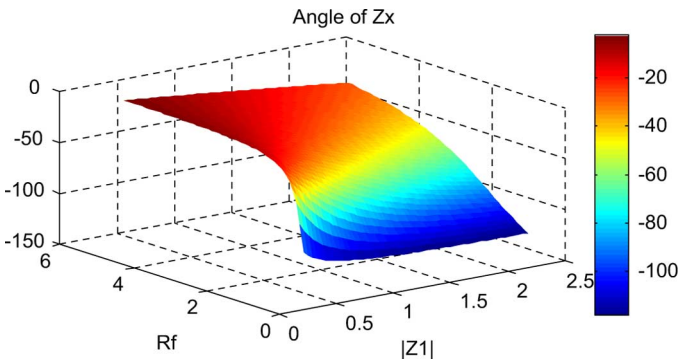


Fig. 2. Z_x with different R_f and $|Z_{b1}|$, $\angle Z_{b1} = 60°$, $Z_{L2} \rightarrow inf.$, $Z_{L\Sigma} = 500\angle 30°$.

Without loss of generality, a fault is assumed at node i with a fault resistance of R_f. V_i can be represented by I_i and R_f

$$\begin{cases} V_i = Z_{\text{equ}} \cdot I_i \\ Z_{\text{equ}} = R_f \| Z_{L1} \| [Z_{b1} + (Z_{L2} \| Z_{L\Sigma})]. \end{cases} \tag{8}$$

Now consider the situation that the fault-location software put "fault" at node $i+1$. The process of postfault load flow is equal to that of putting an impedance of Z_x at node $i+1$ and tuning it to get the same Z_{equ}:

$$Z_{\text{equ}} = Z_{L1} \| [Z_{b1} + (Z_{L2} \| Z_{L\Sigma} \| Z_x)] \tag{9}$$

which yields

$$Z_x = -\frac{Z_{b1} - Z_{\text{equ1}}}{Z_{b1} + Z_{L2}Z_{L\Sigma} - Z_{\text{equ1}}} \cdot Z_{L2}Z_{L\Sigma} \tag{10}$$

where

$$Z_{\text{equ1}} = \frac{R_f + Z_{b1} + \frac{Z_{L2}Z_{L\Sigma}}{Z_{L2}Z_{L\Sigma}}}{R_f \cdot \left(Z_{b1} + \frac{Z_{L2}Z_{L\Sigma}}{Z_{L2}Z_{L\Sigma}}\right)}. \tag{11}$$

When $R_f = 0$, we have $Z_x = -Z_{b1}Z_{L2}Z_{L\Sigma}/Z_{b1} + Z_{L2}Z_{L\Sigma}$. Assuming load impedances to be high enough to be neglected and applying $Z_{b1}/(Z_{L2}Z_{L\Sigma}) \approx 0$, we have $Z_x \approx -Z_{b1}$. Fig. 2 shows the angle and amplitude of Z_{x} with different R_f and $|Z_{b1}|$ when $\angle Z_{b1} = 60°$. It can be seen that although $\angle Z_{\text{x}}$ changes significantly with different settings of R_f and $|Z_{b1}|$, the angle is always negative (impedance vector in the 3rd and 4th quadrant). Similar analysis has been performed on cases where Z_{x} is connected to nodes before node i, and the conclusion is that the angle of Z_{x} is closest to 0° when it is connected to the actual location of fault.

The aforementioned discussion reveals that Pereira's algorithm is not capable of differentiating neighboring or serial nodes in some cases because representation of Z_{x} is not considered.

III. Proposed Fault-Location Method

A. Description of Procedures

The proposed fault-location approach utilizes voltage and current phasors from the root of a feeder and the magnitude of voltage sags from sparse sensors with voltage measurements, such as power-quality meters. Synchronization or phasor-angle information is not required. The feeder is divided into several segments based on the placement of protective devices.

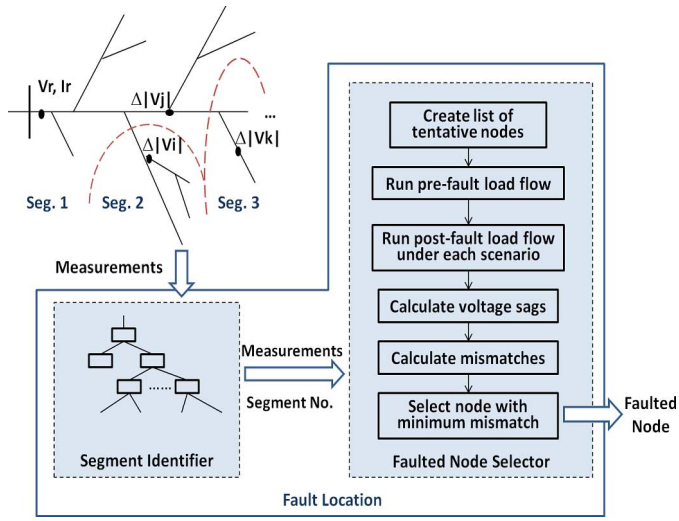


Fig. 3. Procedures of the proposed fault-location scheme.

The proposed fault-location scheme is illustrated in Fig. 3. The upper left of the figure is a diagram of a distribution feeder with segmentation and location of measurements. At the beginning of fault-location process, DT-based segment identifier receives the measurements and identifies the faulted segment. The segment information is then passed on to the function block of faulted node selector, where fault is simulated at every node in the identified segment, and the scenario producing the smallest difference between simulated and measured quantities is selected as the output.

B. DT-Based Segment Identifier

In classification analysis, a case consists of instance (x, y) where x is the vector of predictor variables and y is the target categorical variable. A classification function is used to express the relationship between x and y, through which it is possible to estimate how y changes when x is varied. In our proposed approach, such classification function is realized by a binary tree structure, where x is the vector of measurements used for fault location and y is the fault segment ID.

In this work, the commercial data mining software CART [23] is used to develop the classification trees. The approach in CART to build a DT entails three steps: 1) tree growing using a learning dataset; 2) tree pruning using cross-validation or an independent validation dataset; and 3) selection of the optimal pruned tree. The DT growing, node splitting, tree pruning and optimal tree selection algorithms are detailed in [21]. Experimental tests show that there is a trade-off between DT complexity and its accuracy: a small-sized tree may not be able to capture sufficient system behavior, and a large-sized tree may lead to imprecise prediction due to its over-fitting model. In this work the rule of minimum cost regardless of size to search for the best pruned DT commensurate with accuracy is adopted [24]. The complexity cost parameter in CART is set to zero.

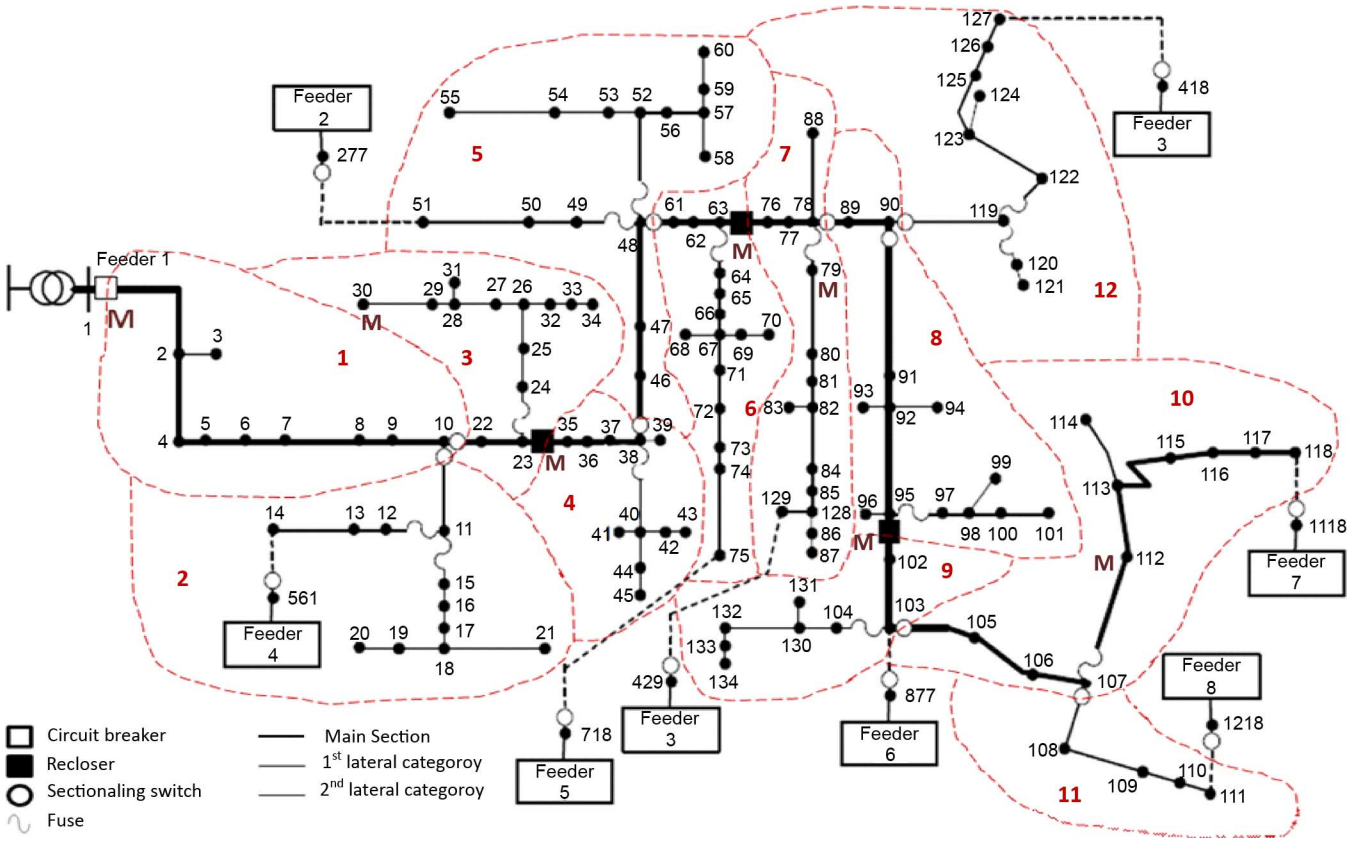


Fig. 4. Topology of the 13.8-kV, 134-node overhead distribution system.

C. Faulted Node Selector

Based on the conclusion from Section II a new criterion for selecting the faulted node is proposed

$$error_j = \frac{\sum_{i=1}^{m} \left| \Delta V_{M,i}^{\text{meas,j}} - \Delta V_{M.i}^{\text{calc,j}} \right|}{V_N} + \alpha \frac{\left| \angle Z_x^j \right|}{2\pi} \qquad (12)$$

where

error_j	mismatch associated with node j assumed as the faulted node;
$\Delta V_{M,i}^{\text{meas,j}}$	measured voltage-sag amplitude at the ith measurement node;
$\Delta V_{M,i}^{\text{calc,j}}$	calculated voltage-sag amplitude at the ith measurement node;
V_N	rated voltage;
α	weight factor for angle index;
$\angle Z_x^j$	angle index in radius.

calculated from

$$\angle Z_x^j = \angle V_j^{\text{calc,j}} - \angle I_f^{\text{calc,j}}. \qquad (13)$$

$\angle V_j^{\text{calc,j}}$ and $\angle I_j^{\text{calc,j}}$ are the calculated angle of node voltage and fault current at node j.

Node with the smallest value of error_j will be selected as the algorithm output. The optimal value of α from (12) is highly dependent on the accuracy of input measurements and the system model. Typically, if the model parameters are close to actual values from the field and the number of voltage measurements is small, or the voltage measurements contain high level of error, a larger weight factor should be assigned to the angle index. When the measurements are accurate but a simplified model is used, smaller value of α will produce better result.

IV. Implementation of the Proposed Method

A. Test System

The proposed fault-location method has been implemented on a 13.8-kV, 134-node, overhead three-phase primary distribution feeder shown in Fig. 4. This is a practical system extracted from the Brazilian distribution network [25]. The total connected load of Feeder 1 is 695.23 MW, and the length of the main section of the feeder is 432 km. Total length of first and second category laterals is 267 km and 261 km respectively. The average distance between two neighboring nodes (load taps) is 7.2 km. The maximum and minimum distances between neighboring nodes are 90 km and 1 km, respectively.

A nontransposed line model with lumped parameters were used, and loads were modeled as constant impedances in the Alternative Transients Program (ATP) [26] simulations. Root voltage and current are measured at node 1. Six voltage measurements are placed along the feeder, at nodes 23, 30, 63, 79,

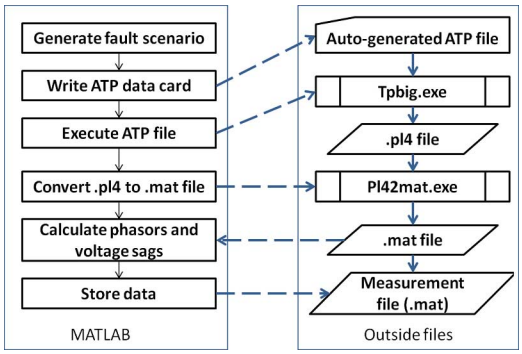


Fig. 5. Procedure of knowledge base generation.

96, and 112, respectively (marked as M in Fig. 4). The feeder is divided into 12 segments based on the placement of reclosers and sectionalizing switches (numbered with dotted curves in Fig. 4).

B. Generation of Knowledge Base

The knowledge base is a database used for offline training of the DT-based segment identifier. It is composed of a number of instances, and each instance represents a fault scenario and is labeled with the corresponding fault segment ID. Typically, the DT-based identification model will gain more generalization power if a larger number of instances are included in the knowledge base. However, the database generation process should be properly designed; otherwise, it will not capture sufficient information from the entire problem space.

In this paper, the distribution system shown in Fig. 4 is modeled in ATP. In order to create a sufficiently large knowledge base, add-on scripts for scenario generation have been developed using hybrid programming between MATLAB [27] and ATP. The function takes the original ATP model as a reference model, automatically inserts fault scenario settings into switch and impedance data cards (faulted node, and fault resistance), saves modified model in a separate ATP file and calls execution file "tpbig.exe" to run simulation in ATP. When ATP simulation is complete, the output file from ATP (.pl4 file) is converted to MATLAB data file (.mat file) by calling "pl42mat.exe", the phasors from the feeder root and voltage sags at measurement nodes are calculated in MATLAB and stored with fault information. The process of generating one fault scenario is shown in Fig. 5. The arrows in the left-hand side block illustrate the sequence of MATLAB functions, the arrows in the right-hand side block show the information flow between outside files, and the dashed arrows in between show the calling and returning of outside files.

C. Training of the DT

A knowledge base comprising 49210 fault scenarios is used for DT training. Random errors following a normal distribution with zero mean and deviation of 0.5% are added to the measurements of each scenario to mimic a situation in a real-world. Settings of fault scenarios include fault resistance, faulted node, and prefault load pattern. Faults along the feeder (node 2 to

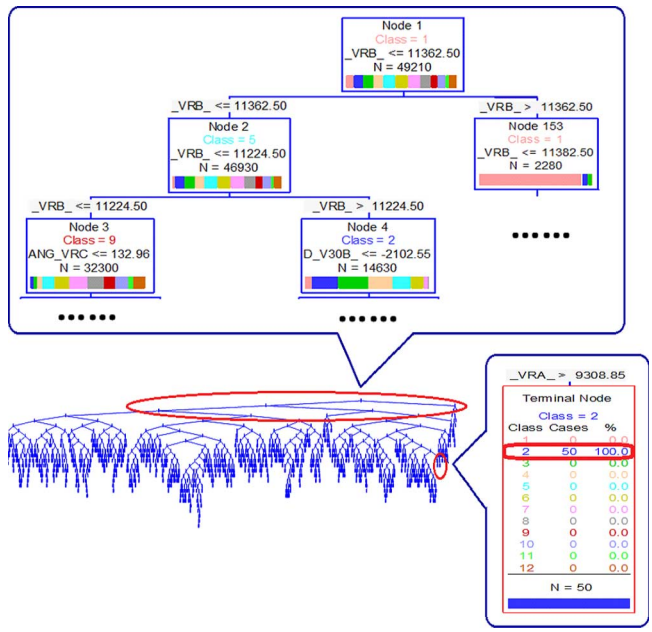


Fig. 6. DT topology for segment identification.

134), with fault resistance of 0 to 30 Ω are simulated. Fault types are predetermined by the change in phase voltage amplitude, phase-to-phase angle and zero-sequence current amplitude. Loads are classified into residential and business, and load variation is achieved by varying the load impedance based on an hourly load forecast of the different types of load.

The 10-fold cross-validation method is used to develop the classification tree in CART. The topology of resulting optimal tree is shown in the middle of Fig. 6. The block above the tree shows details of the four nodes at the top layers. Details of one terminal node are shown in the block on bottom-right. The label of a terminal node is determined by the majority of training cases falling into that node. In this example 50 of the training cases reached the terminal node and they all belong to Class 2. In online applications, the measurements of a fault will be fed into the tree and go through a particular top-down path. Once they reach one terminal node, the faulted segment can be immediately identified.

The computation time for generating fault scenarios depends highly on the number of outputs from ATP simulations. Executed on an Intel Xeon 2.80-GHz CPU with 6 GB of RAM, the average time for completing one scenario on the test model with six voltage outputs and one current output, is about 3 s. However, in the study of optimal sensor placement in Section VI, generating a scenario with 21 voltage outputs and 1 current output takes about 10 s. The time for DT training is much shorter. It takes less than 2 min to grow, prune, and select the best pruned DT for the examined 134-node feeder network. The computation time is estimated using the built-in clock of MATLAB and CART.

To embed the segment identifier in online applications, a unique DT should be developed for each network, since for different feeder configurations, different knowledge bases need to be formulated.

TABLE I
DESCRIPTION OF SCENARIO GROUPS AND RATE OF SUCCESSFUL SEGMENT IDENTIFICATION

Scenario Group	Description	Accuracy of Segment Identification (%)
1	R_f=0Ω, no measurement error	**100**
2	R_f=1Ω, no measurement error	**98.5**
3	R_f=5Ω, no measurement error	**100**
4	R_f=0Ω, σ(error)=0.5%	**94.7**
5	R_f=1Ω, σ(error)=0.5%	**92.5**
6	R_f=5Ω, σ(error)=0.5%	**91.7**
7	R_f=1Ω, load at node 21 - 60 increased by 100%	**98.4**
8	R_f=1Ω, load at node 21 - 60 decreased by 50%	**99.7**
9	R_f=1Ω, load power factor at node 21 - 60 shifted from 0.92 to 0.85	**95.5**

D. Implementation of the Faulted Node Selector

MATLAB programs are developed to realize the node selection algorithm. The optimal weight factor of α is determined by the following procedures: 1) vary α in the range of 0 to 0.1; 2) feed the fault-location program with no-error measurements and record the output error; 3) fit the sets of α and output errors to a polynomial curve; and 4) find the extreme point on the curve and record α. The optimal α is determined as 0.031. Both the algorithm reported in [15] and the proposed algorithm are implemented, and the results will be compared in Section V.

V. CASE STUDIES

A. Overview of Case Studies

To examine the performance of the proposed method, 1197 fault scenarios and corresponding measurements have been generated as the test cases. None of these scenarios were used during the DT training phase. The generated fault scenarios belong to nine groups. In each group, 133 fault scenarios corresponding to the faults occurring at nodes 2 to 134 were simulated. The detailed description of each scenario group is provided in Table I.

B. Performance of the DT-Based Segment Identifier

With the offline training described in Section IV-C, the success rates of the DT-based segment identification are also reported in Table I. An initial observation of test results reveals that the segment identifier is capable of maintaining a success rate of above 98.5% in all three scenario groups where measurements are assumed errorless. For Scenario Group 4 to 6, in which the measurement errors were considered, the prediction accuracy reduced a little bit, and accuracy greater than 91% were reached for all three groups. In Scenario Group 7 to 9, the loads from node 21 to node 60 were varied and the DT performance was tested. As shown in the table, identification accuracy higher than 95.5% was achieved for each group.

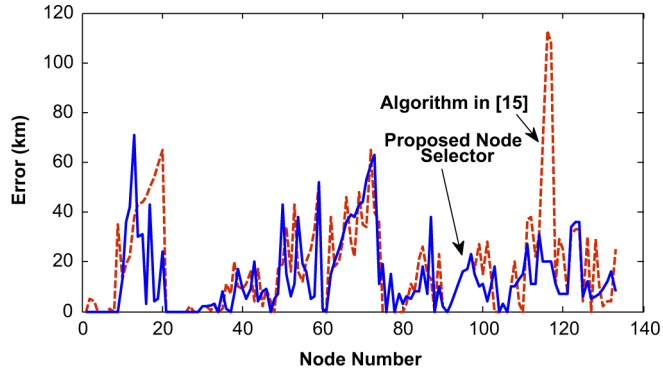


Fig. 7. Fault-location errors in kilometers with faults along the feeder.

In the preparation of the knowledge base, two simulation steps were utilized: 1) from $R_f = 0\ \Omega$ to $R_f = 25\ \Omega$, with step of $0.1\ \Omega$; 2) from $R_f = 25\ \Omega$ to $R_f = 70\ \Omega$, with step of $3\ \Omega$. In Table I the results of fault resistance up to 5 ohm were reported. The DT performance for the other scenarios, where the fault resistance is larger than $3\ \Omega$, was also evaluated. There was a drop of prediction accuracy when fault resistance is larger than $25\ \Omega$. This is because the training cases around those resistance values are not as adequate as the cases for a resistance smaller than $25\ \Omega$. The problem of fault segment identification is nonlinear. The more system behavior captured in the knowledge base, the better the DT will be trained, and therefore higher prediction accuracy will be achieved when it is embedded online.

C. Performance Under Perfect Condition

Scenario Groups 1 to 3 are used for tests under "perfect condition". The load information given to the fault-location program is consistent with the settings of load impedances in ATP and the measurement values are considered accurate. Under such condition the error in fault location comes from the simplification of line model (shunt capacitor being neglected) and computation error.

1) Comparison Before Introducing Segment Identification: Fig. 7 shows the comparison of the method from [15] and the proposed node selection method (without segment identification) for Scenario Group 2. The x axis shows the faulted node number, and the y axis is the output error represented by the distance between calculated and actual location of faults in kilometers. The dotted curve is the error from Pereira's method, and the solid one is the error from the proposed method in Section III-C.

On average, the proposed node selector reduces the errors by 34.1%. The mean of errors with faults on the feeder main section has dropped from 15.3 to 6.15 km, which is less than the average distance between two neighboring nodes. Although, in general, both methods show better performance with faults on the main section of feeder, the performance goes down as faults occur on nodes toward the end of laterals, for example, nodes 14 and 75. At node 116, Pereira's method selected node 127, causing the error to be higher than 110 km, but the proposed method avoided this error.

The main window in Fig. 8 illustrates a successful node selection for one of the test fault scenarios. The smallest mismatch

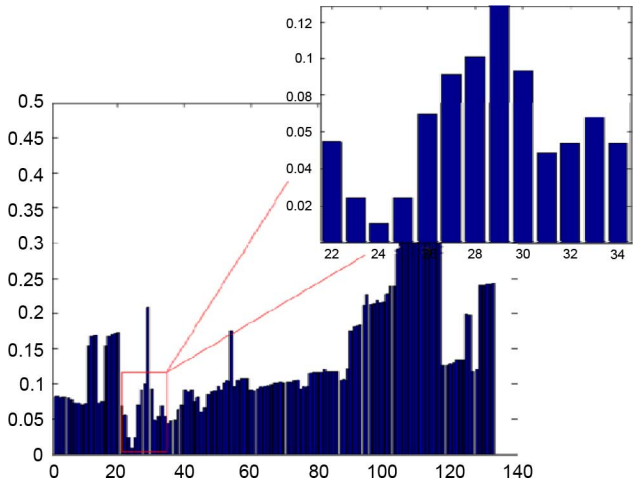


Fig. 8. Mismatch calculated for the fault at node 24 $R_f = 0$.

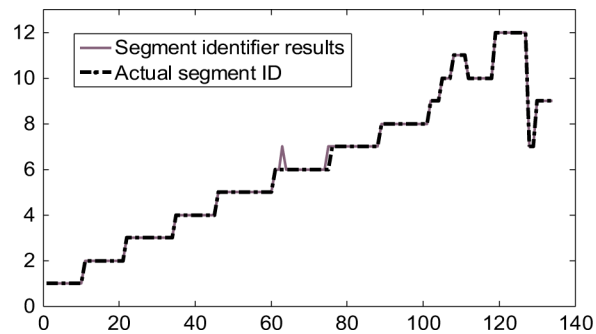


Fig. 9. Output from the segment identifier, with $R_f = 1\ \Omega$.

is observed at node 24, which is indeed the actual location of fault.

2) Further Improvement With Segment Identification: Fig. 9 shows the outputs from the segment identifier with faults at nodes 2 to 134, $R_f = 1\ \Omega$. The solid line is the actual segment number, and the dashed one shows the segment number identified by the DT. In this group of fault scenarios, nodes 63 and 75 are misclassified.

Fig. 10 shows the reduction of error by introducing the segment identifier. The dotted line represents errors from the method by only using the node selector only (solid line in Fig. 9) and the solid line shows errors after utilizing the segment identifier. Segment IDs from the DT tree have a success rate of 99.7% for $0\ \Omega$ faults, 98.5% for $1\ \Omega$ faults, and 100% for $5\ \Omega$ faults. It can be seen that spikes at nodes 74, 87, 101, and some other nodes have been alleviated because the node outside the selected segment has been removed from the list of tentative nodes. However, the error at node 75 did go up because the node has been misclassified into Segment 7.

The computational burden is reduced significantly. For example, both methods are able to successfully locate the fault at node 24 (Fig. 8). Instead of running load flow for the fault being at nodes 2 to 134, the proposed method takes nodes 22 to 34 as tentative nodes and performs load-flow calculation, which reduced the computation by nearly 90%. This means only the nodes in the zoom-in window of Fig. 8 were investigated in the proposed algorithm. In the meantime, the time for performing

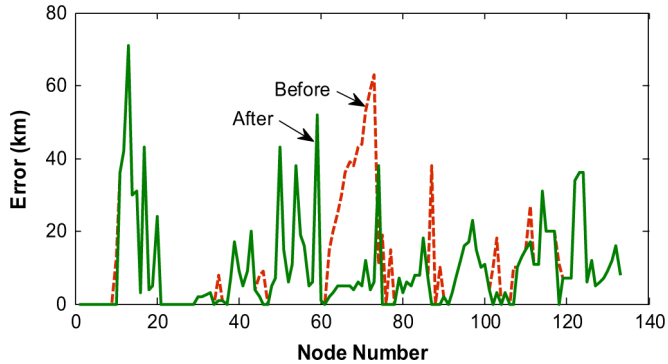


Fig. 10. Improvement with the segment identifier.

TABLE II
ERRORS IN PERFECT CONDITION

Rf (Ω)	Total		Main		L1		L2	
	Alg1	Alg2	Alg1	Alg2	Alg1	Alg2	Alg1	Alg2
0	17.8	11.1	13.6	7.10	17.4	13.4	21.5	12.3
1	18.2	9.49	15.3	3.68	19.6	13.6	19.3	10.6
5	24.7	17.3	20.3	8.90	29.3	19.7	24.3	21.9

segment identification using a properly trained DT on a scenario is negligible compared to that for fault-location calculation.

3) Impact of Fault Resistance: The mean of errors from different settings of fault resistance are recorded in Table II. "Main," "L1, "L2" refers to scenarios with a fault on main section I 1st category laterals and 2nd category laterals, respectively; "Alg1" and "Alg2" represent algorithms from [15] and the one proposed in this paper, respectively. The comparison clearly reveals better performance of the proposed algorithm. Although theoretically the proposed method should not be affected by fault resistance, the test results show otherwise. The accuracy from the node selector gradually decreases as the fault resistance goes up. This is because when fault resistance is high, the differences between voltage sags are reduced, and their dominance over the calculated mismatch is compromised by computational errors. Nevertheless, the proposed algorithm constantly produces superior results and shows a slower deterioration of accuracy over increasing fault resistance.

D. Performance Under a Nonperfect Condition

The impact of measurement error and inconsistent load condition are evaluated in the test of nonperfect condition (Scenario Groups 4 to 9).

1) Impact of Measurement Error: Scenario Groups 4 to 6 are designed to evaluate the impact of measurement error. Random values of error with a mean of 0 and standard deviation of 0.5% of rated voltage are added to the measurements. The results are recorded in rows 1 to 3 of Table III.

2) Impact of Load Condition: Scenarios for evaluating the impact of load condition are generated by varying the load impedance in the ATP model, without updating the load profile used by the fault-location program. Loads are varied as described in Table I, Scenario Group 7 to 9. Fault resistance is set as $1\ \Omega$. Results are recorded in rows 4 to 6 of Table III. The

TABLE III
TEST RESULT FROM NONPERFECT CONDITION SCENERIOS

Scn. Grp.	Total		Main		L1		L2	
	Alg1	Alg2	Alg1	Alg2	Alg1	Alg2	Alg1	Alg2
4	21.9	11.8	11.4	6.83	24.9	16.0	27.8	12.2
5	26.7	15.4	31.3	5.70	24.7	21.5	24.6	17.8
6	35.6	22.2	27.6	18.5	37.1	22.2	40.7	25.2
7	20.9	12.6	16.6	12.8	25.2	12.9	20.6	12.1
8	18.7	8.73	19.6	3.15	17.6	12.3	19.0	10.2
9	19.9	10.4	15.2	4.5	23.6	15.8	20.4	10.5

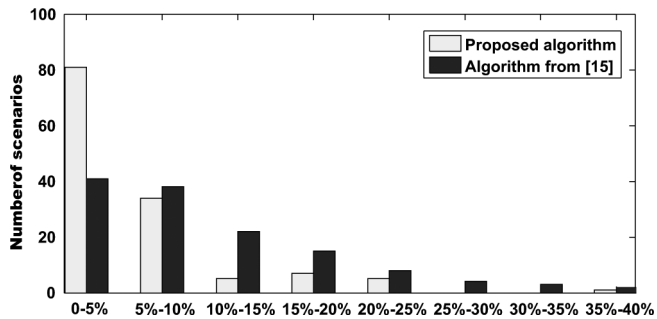


Fig. 11. Errors under load variation (scenario group 8).

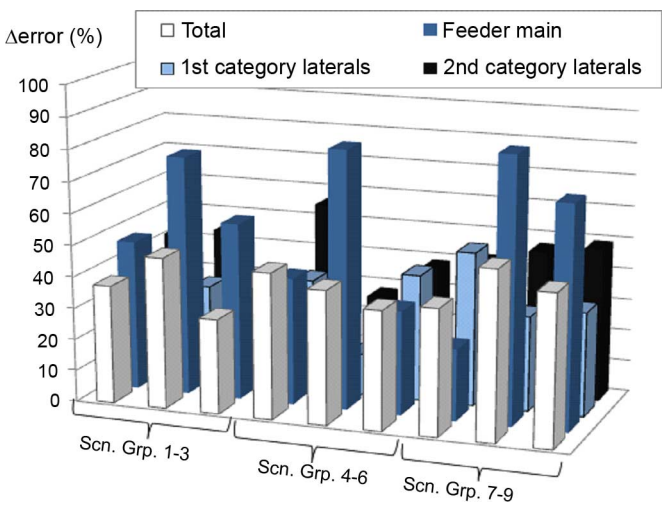


Fig. 12. Reduction in fault-location errors.

histogram of the proportions of errors to the distances of actual location from feeder root for fault scenarios in Scenario Group 8 is shown in Fig. 11. Most of the results from the proposed algorithm contain an error of less than 10% of the distance to fault, while Pereira's algorithm produces more results with larger errors.

Fig. 12 shows the percentage of reduced errors from nine scenario groups. Generally, the errors with faults on the main section of the feeder are reduced most significantly, with the highest being more than 80%. In every scenario group, the mean error for each line type has been reduced.

3) Impact of Missing Data: A practical concern that almost every fault-location method needs to deal with is the data missing due to communication errors or failed sensors. One major advantage of the proposed approach, compared to the conventional methods, is that the DT has the capability to automatically deploy a backup measurement when the primary measurement is lost. Backup measurements, called *Surrogates* in DT, are highly correlated with the primary splitters, contain similar information, and have almost identical power to split a tree node. During online application, once the variable that previously split a tree node is missing, its surrogate will serve as the primary splitter without a significant degradation in the overall accuracy of the fault-location algorithm.

In the meantime, the proposed algorithm for selecting the faulted node has a flexible number of voltage sag inputs. This means, although more voltage measurements from the system suggest better prediction of fault location, the algorithm is able to produce satisfactory results once one or two measurements are missing. To support the statement, the measurement on node 63 was removed from the testing cases and scenarios from Group 6 were repeated. The DT identified 90.2% of the segments correctly. The mean error from the proposed method is 23.9 km, which is higher than that from the fault location using six voltage-sag inputs. Yet, the error is still lower than the results from Alg. 1 with no missing data.

VI. OPTIMAL SENSOR PLACEMENT

While the proposed algorithm will most likely achieve the best fault-location result by assuming the measurement units are installed at every feeder node, it is not economically feasible in practice to do so due to high expenses of the corresponding communication paths as well as the sensors themselves. A reasonable approach may be to install only a limited number of sensors at the most critical feeder nodes. Conventionally, the locations of measurement units are determined using engineering insight and empirical evidence. Recently, the concept of observability from state estimation has been borrowed for fault-location applications [28], [29]. In this paper, a different approach will be deployed to find the best sensor locations in a quantitative way.

A. Feature Selection Using Cart

The problem of finding the optimal sensor locations is equivalent to selecting the best reduced set of DT input variables given a pool of candidate measurements. Ideally, the optimal solution could be obtained through an exhaustive trial and comparison of all possible combinations. However, it is computationally too involved to do so. The feature selection property of DT has been explored in [30] to derive a reduced input dataset. In this paper, it has been extended to distribution systems to quantitatively measure the importance of feeder nodes in fault-location applications.

A close observation of the DT model structure shown in Fig. 6 reveals that each tree node is split by an input variable. The variable is determined by searching all candidate predictors, and finding the split which gives the largest decrease in class impurity. The variables gain credit toward their contribution by serving as primary splitters that actually split a node, or as backup splitters (surrogates) to be used when the primary splitter is missing. By summarizing the variables' contribution to the overall tree when all nodes are examined, the *variable importance (VI)* is obtained.

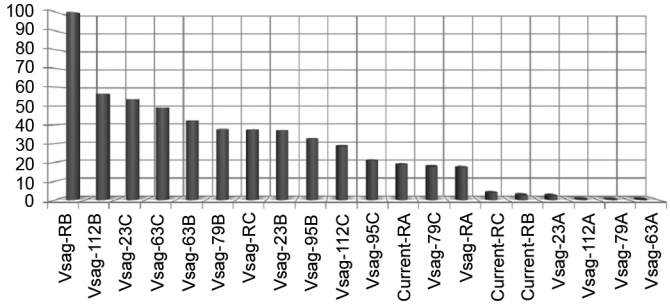


Fig. 13. Variable importance for fault segment identification.

To calculate the *VI*, search all candidate splits $s \in S$ at each tree node $t \in T$, and find the split $\tilde{s}_m$ which gives the largest decrease in impurity I [21]

$$\Delta I(\tilde{s}_m, t) = \max_{s \in S} \Delta I(s, t)\,. \tag{14}$$

The measure of importance of variable x_m is defined as

$$VI(x_m) = \sum_{t \in T} \Delta I(\tilde{s}_m, t). \tag{15}$$

Fig. 13 shows the variable importance derived in Section V-B. It can be easily observed that several measurements (e.g., *Vsag-RB*) (phase B voltage sag at feeder root) and *Vsag-112B* (phase B voltage sag at node 112), have much higher importance compared to some other variables, such as *Vsag-79A* and *Vsag-63A*.

B. Optimal Sensor Placement

In brief, the idea of optimal sensor placement is: for each feeder node i, its overall contribution to the fault segment identification can be quantified by combining the importance of variables measured at node i.

The *Node importance (NI)* is defined to quantitatively measure the contribution of each feeder node to fault segment identification, and mathematically it can be expressed as

$$NI_i = \sum_{v \in V} VI(v, i \in v) \tag{16}$$

where V is the set of DT input variables, v is the individual variable belonging to V, and *VI* is its variable importance. By specifying $i \in v$, only the variables measured at node i will be counted.

The *NI* reflects the contribution of each node to fault segment identification. The higher the *NI*, the more important the feeder node is to the proposed algorithm. Therefore, the optimal sensor locations are suggested by selecting the top-ranked nodes. In this paper, the *NI* of top-ranked nodes is computed by considering only the primary splitters, because the surrogate variables that appear to be important but rarely split tree nodes are almost certainly highly correlated with the primary splitters and contain similar information. Once the top-ranked nodes are selected, the standard variable importance considering both primary and surrogate splitters is used to rank the remaining nodes. A set of 21 nodes from the examined distribution system is first selected as candidates using engineering judgment, most of which are intersections along the main feeder. In Table IV, the *NI* for the 21 candidates is calculated and the top eight nodes are listed. Also shown in the table are the eight nodes with the lowest *NI*. In practice, the voltage and current measurements are usually available from feeder root; therefore, in the following discussion, it is assumed that one sensor is installed at the feeder root.

TABLE IV
BUS IMPORTANCE RANKING OF THE FEEDER SYSTEM

Top Ranked Feeder Nodes			Lowest Ranked Feeder Nodes		
Rank	Location	NI	Rank	Location	NI
# 1	**Root**	**226.52**	**# 14**	**Node 67**	**49.31**
# 2	**Node 107**	**138.68**	**# 15**	**Node 127**	**46.87**
# 3	**Node 23**	**124.21**	**# 16**	**Node 60**	**43.49**
# 4	**Node 103**	**113.34**	**# 17**	**Node 38**	**43.48**
# 5	**Node 14**	**108.90**	**# 18**	**Node 90**	**41.96**
# 6	**Node 63**	**105.87**	**# 19**	**Node 48**	**41.37**
# 7	**Node 123**	**95.08**	**# 20**	**Node 82**	**39.97**
# 8	**Node 18**	**83.46**	**# 21**	**Node 52**	**37.88**

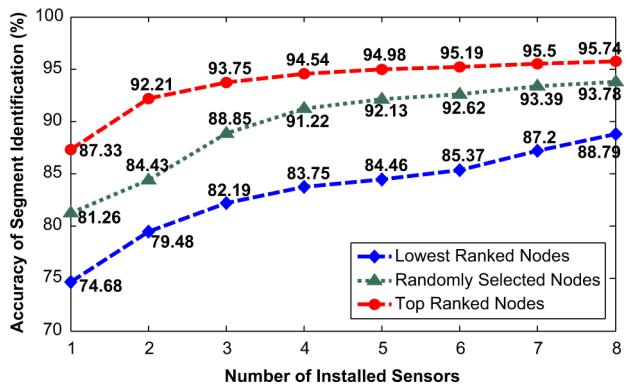


Fig. 14. DT performance considering different sensor placement.

C. Fault-Location Accuracy

1) Segment Identification: Suppose that apart from the feeder root, another 1 to 8 sensors are planned for installation in the feeder network. By placing them at the top-ranked candidate nodes of Table IV, and considering measurement errors, the resulting accuracy in segment identification for the case of $\mathrm{R_f} = 5\ \Omega$ is summarized in Fig. 14. The DT performances using the measurements from the lowest ranked nodes and from randomly selected nodes are also shown in the figure, respectively, for the purpose of comparison. For the case of randomly selected nodes, the process has been replicated until the mean and standard deviation of DT accuracy become stable.

An observation from Fig. 14: in contrast with the DTs fed with measurements from the lowest ranked feeder nodes or randomly selected nodes, the DTs constructed using the measurements from the top-ranked nodes have achieved better accuracy in segment identification.

2) Fault Node Selection: Since the fault-location algorithm takes the same input measurements as the segment identifier, the optimal sensor placement is expected to have a positive impact on it as well.

The fault node selection algorithm was executed with fault scenarios from Group 6, using measurements from the six top-ranked nodes (Node 107, 23, 103, 14, 63, and 123). The procedure was then repeated using measurements from the six lowest ranked nodes (Node 52, 82, 48, 90, 38, and 60). The resulting fault-location mean errors are 19.7 km and 25.4 km, respectively. Last but not least, the measurements from the original five sensor locations shown in Fig. 4 (Node 23, 63, 79, 95, and 112) plus one more measurement point at Node 119 were utilized. The resulting fault-location mean error is 20.9 km. The proposed optimal sensor placement methodology has exhibited encouraging capability for improving fault-location estimation.

VII. Conclusion

This paper proposes an algorithm for automated fault location in radial distribution systems. The following conclusions have been reached:

- The computational complexity of voltage-sag-based fault-location algorithms has been significantly reduced by utilizing the DTs for fault segment identification.
- A new algorithm for faulted node selection has been proposed and proven to be more accurate theoretically and experimentally.
- The proposed method has been implemented on an actual distribution system. Experimental analysis indicates better performance of fault-location accuracy and reliability.
- The algorithm has been tested extensively under different simulation scenarios. The results show that the proposed method is able to handle a certain degree of measurement error and load variations.
- The DT variable importance was used to suggest optimal sensor placement. Test results show that the measurements from suggested feeder nodes lead to a higher fault segment identification and fault-location accuracy.

Acknowledgment

The authors of this paper would like to thank Dr. R. A. Fernandes Perreira for providing the practical system model used in this paper during his stay at Texas A&M University as a Visiting Scholar.

References

[1] J. Northcote-Green and R. Wilson, *Control and Automation of Electrical Power Distribution Systems*. New York: Taylor & Francis, 2006.

[2] R. Horn and P. Johnson, "Outage management applications and methods panel session: Outage management techniques and experience," in *Proc. IEEE Power Eng. Soc. Winter Meeting*, Feb. 1999, vol. 2, pp. 866–869.

[3] S. T. Mak, "A synergistic approach to using AMR and intelligent electronic devices to determine outages in a distribution network," presented at the Power Syst. Conf., Clemson, SC, USA, 2006.

[4] K. Sridharan and N. N. Schulz, "Outage management through AMR systems using an intelligent data filter," *IEEE Trans. Power Del.*, vol. 16, no. 4, pp. 669–675, Oct. 2001.

[5] A. A. Girgis, C. M. Fallon, and D. L. Lubkeman, "A fault location technique for rural distribution feeders," *IEEE Trans. Ind. Appl.*, vol. 29, no. 6, pp. 1170–1175, Dec. 1993.

[6] R. Das, "Determining the locations of faults in distribution systems," Ph.D. dissertation, Saskatchewan Univ., Saskatoon, SK, Canada, 1998.

[7] L. Yuan, "Generalized fault-location methods for overhead electric distribution systems," *IEEE Trans. Power Del.*, vol. 26, no. 1, pp. 53–64, Jan. 2011.

[8] S. Das, N. Karnik, and S. Santoso, "Distribution fault-locating algorithms using current only," *IEEE Trans. Power Del.*, vol. 27, no. 3, pp. 1144–1153, Jul. 2012.

[9] R. H. Salim, M. Resener, A. D. Filomena, K. R. Caino de Oliveira, and A. S. Bretas, "Extended fault-location formulation for power distribution systems," *IEEE Trans. Power Del.*, vol. 24, no. 2, pp. 508–516, Apr. 2009.

[10] M. S. Choi, S. J. Lee, D. S. Lee, and B. G. Jin, "A new fault location algorithm using direct circuit analysis for distribution systems," *IEEE Trans. Power Del.*, vol. 19, no. 1, pp. 35–41, Jan. 2004.

[11] A. Borghetti, M. Bosetti, C. A. Nucci, M. Paolone, and A. Abur, "Integrated use of time-frequency wavelet decompositions for fault location in distribution networks: Theory and experimental validation," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 3139–3146, Oct. 2010.

[12] A. Borghetti, M. Bosetti, M. D. Silvestro, C. A. Nucci, and M. Paolone, "Continuous-wavelet transform for fault location in distribution power networks: Definition of mother wavelets inferred from fault originated transients," *IEEE Trans. Power Syst.*, vol. 23, no. 2, pp. 380–388, May 2008.

[13] F. H. Magnago and A. Abur, "Fault location using wavelets," *IEEE Trans. Power Del.*, vol. 13, no. 4, pp. 1475–1480, Oct. 1998.

[14] Z. Galijasevic and A. Abur, "Fault location using voltage measurements," *IEEE Trans. Power Del.*, vol. 17, no. 2, pp. 441–445, Apr. 2001.

[15] R. A. F. Pereira, L. G. W. Silva, M. Kezunovic, and J. R. S. Mantovani, "Improved fault location on distribution feeders based on matching during-fault voltage sags," *IEEE Trans. Power Del.*, vol. 24, no. 2, pp. 852–862, Apr. 2009.

[16] S. Lotfifard, M. Kezunovic, and M. J. Mousavi, "Voltage sag data utilization for distribution fault location," *IEEE Trans. Power Del.*, vol. 26, no. 2, pp. 1239–1246, Apr. 2011.

[17] M. Togami, N. Abe, T. Kitahashi, and H. Ogawa, "On the application of a machine learning technique to fault diagnosis of power distribution lines," *IEEE Trans. Power Del.*, vol. 10, no. 4, pp. 1927–1936, Oct. 1995.

[18] Y. Sheng and S. M. Rovnyak, "Decision tree-based methodology for high impedance fault detection," *IEEE Trans. Power Del.*, vol. 19, no. 2, pp. 533–536, Apr. 2004.

[19] H. T. Yang, W. Y. Chang, and C. L. Huang, "Power system distributed on-line fault section estimation using decision tree based neural nets approach," *IEEE Trans. Power Del.*, vol. 10, no. 1, pp. 540–546, Jan. 2004.

[20] S. R. Samantaray, "Decision tree-based fault zone identification and fault classification in flexible ac transmissions-baesd transmission line," *IET Gen. Transm. Distrib.*, vol. 3, no. 5, pp. 425–436, 2009.

[21] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth, 1984.

[22] C. S. Cheng and D. Shirmohanmmadi, "A three-phase power flow method for real-time distribution system analysis," *IEEE Trans. Power Syst.*, vol. 10, no. 2, pp. 671–679, May 1995.

[23] D. Steinberg and G. Mikhail, *CART 6.0 User's Manual*. San Diego, CA: Salford Systems, 2006.

[24] C. Zheng, V. Malbasa, and M. Kezunovic, "A fast stability analysis scheme based on classification and regression tree," presented at the IEEE Conf. Power Syst. Technol., Auckland, New Zealand, Oct. 2012.

[25] A. A. P. Biscaro, R. A. F. Pereira, and J. R. S. Mantovani, "Optimal phasor measurement units placement for fault location on overhead electric power distribution feeders," in *Proc. IEEE Power Energy Soc. Transm. Distrib. Conf. Expo. Latin America*, 2010, pp. 37–43.

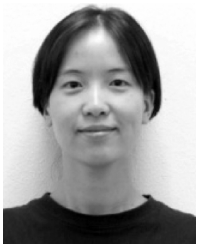
[26] *"ATP/EMTP Rule Book,"* Argentinian EMTP/ATP User Group, Argentina, 2002.

[27] *"MATLAB R2012b User's Guide,"* Mathworks Inc., Nattick, MA, USA. [Online]. Available: http://www.mathworks.com

[28] K. P. Lien, C. W. Liu, C. S. Yu, and J. A. Jiang, "Transmission network fault location observability with minimal pmu placement," *IEEE Trans. Power Del.*, vol. 21, no. 3, pp. 1128–1136, Jul. 2006.

[29] M. Korkali and A. Abur, "Optimal deployment of wide-area synchronized measurements for fault-location observability," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 482–489, Feb. 2012.

[30] C. Zheng, V. Malbasa, and M. Kezunovic, "Regression tree for stability margin prediction using synchrophasor measurements," *IEEE Trans. Power Syst.*, to be published.



Yimai Dong (S'07) received the B.S. and M.S. degrees in electrical engineering from North China Electric Power University, Beijing, China, in 2005 and 2007 respectively, and is currently pursuing the Ph.D. degree in electrical engineering at Texas A&M University, College Station, TX.

Her research interests include power system optimization, fault location, distribution outage management, and reliability assessment.

Ce Zheng (S'07) received the B.S. and M.S. degrees in electrical engineering from North China Electric Power University, Beijing, China, in 2005 and 2007, respectively, and is currently pursuing the Ph.D. degree in electrical engineering at Texas A&M University, College Station.

His research interests include data-mining techniques applied to power system stability analysis, synchrophasor applications, and the impact analysis of grid integration of distributed generation.

Mladen Kezunovic (S'77–M'80–SM'85–F'99) received the Dipl. Ing. degree in electrical engineering from the University of Sarajevo in 1974, and the M.S. and Ph.D. degrees in electrical engineering from the University of Kansas in 1977 and 1980, respectively.

Currently, he is the Eugene E. Webb Professor, Director of the Smart Grid Center, Site Director of National Science Foundation (NSF) I/UCRC "Power Engineering Research Center, PSerc"; and Deputy Director of another NSF I/UCRC "Electrical Vehicles: Transportation and Electricity Convergence, EV-TEC." He has published more than 450 papers, given over 100 seminars, invited lectures and short courses, and consulted for over 50 companies worldwide. He is the Principal of XpertPower Associates, a consulting firm specializing in power systems data analytics. His main research interests are digital simulators and simulation methods for relay testing, as well as the application of intelligent methods for power system monitoring, control, and protection.

Dr. Kezunovic is a member of CIGRE and a Registered Professional Engineer in Texas.