ABSTRACT

CAI, YIXIN. Small World Stratification for Power System Fault Diagnosis. (Under the direction of Dr. Mo-Yuen Chow.)

Power system fault diagnosis aims to expedite the repair process after a power outage by providing information about the root cause. In distribution systems, automated fault diagnosis algorithms learn the relationship between fault root causes and environmental factors from historical fault events and infer the root cause of those under study. As distribution systems are spatially dispersed and heterogeneous, sampling historical fault events from a small geographic area is preferred in order to focus on the local fault characteristics. However, a small sampling area may not provide enough data for the algorithms to make proper inference.

In this work, the sampling issue in distribution fault diagnosis is studied and the Small World Stratification sampling strategy is proposed. Outage records from the distribution systems of Progress Energy Carolinas Inc. are first integrated with environmental data from other resources and then analyzed to reveal the effect of sampling. To facilitate the study of sampling strategy, a framework for cause-effect modeling and spatial-temporal simulation of fault events is established and a power distribution fault simulator is developed accordingly. The proposed Small World Stratification strategy is demonstrated with simulated fault events and tested with the real-world outage records as well.

Small World Stratification for Power System Fault Diagnosis

by
Yixin Cai

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Electrical Engineering

Raleigh, North Carolina

2011

APPROVED BY:

Dr. Peter Bloomfield	Dr. James J. Brickley
Dr. Richard E. Brown	Dr. Huaiyu Dai

Dr. Mo-Yuen Chow
Chair of Advisory Committee

DEDICATION

To My Parents

Zhihong Cai and Shumin Yuan

To My Wife

Jia Wang

BIOGRAPHY

Yixin Cai was born in Zhengzhou, Henan province, People's Republic of China. He received his Bachelor of Engineering degree in automation in 2003, and his Master of Science degree in control science and engineering in 2006, both from Tsinghua University, Beijing, China. He is currently a Ph.D. candidate at North Carolina State University.

Yixin joined the Advanced Diagnosis, Automation and Control (ADAC) Laboratory under the direction of Dr. Mo-Yuen Chow in fall 2006. His research interests include intelligent fault management and fault diagnosis of power distribution systems, Geographic Information Systems in power systems, and spatial-temporal modeling of power system faults.

ACKNOWLEDGMENTS

This dissertation could not be completed without many helpful suggestions and motivations from several professors, friends, and my family.

First I would like to extend the most sincere gratitude to my advisor, Dr. Mo-Yuen Chow. Dr. Chow has spent his tireless effort and precious time to supervise, encourage, and inspire me on my research. He also provides chances for me to get involved in the professional community, helps me to grow as a young researcher.

Also, I would like to thank Dr. Peter Bloomfield, Dr. James J. Brickley, Dr. Richard E. Brown, and Dr. Huaiyu Dai for being on my advisory committee, Dr. Wenbin Lu and Dr. Lexin Li for their help on statistical feature selection methods and performance evaluation tools, Dr. Simon Hsiang for his comments and inputs on small-world networks, Dr. Le Xu from Quanta Technology for his help on understanding the fault diagnosis problem, Mr. Glenn C. Lampley and Mr. John W. Gajda from Progress Energy Carolinas Inc. for providing data and industrial perspective for my research work.

I would like to thank Mr. Jann Lin for his selfless help in my first two years of life in USA, and all my ADAC lab mates for their help and support.

And last but not least, I would like to thank my mother Shumin Yuan, my father Zhihong Cai, and my wife Jia Wang, for their endless love, support and trust in me all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter VI

Chapter VII

CHAPTER I

INTRODUCTION

I. POWER SYSTEM FAULT DIAGNOSIS

Power systems are vital lifelines of the modern society for maintaining adequate and reliable flows of energy. To cover their service territory, power lines extend thousands of miles, link numerous voltage conversion equipment and protective devices together, and form an interconnected network. As the retail part of power systems, typical distribution systems are geographically dispersed and hence exposed to harsh and uncertain environments. Therefore, the systems can easily be affected by various outage-causing events such as equipment failures, animal contacts, trees, lightning strikes, etc [1].

As electric devices are used almost everywhere, the cost of interruptions to the power supply is increasingly significant. For some industries, the costs can be as high as several million dollars per hour [2]. Accordingly, a fast service restoration of the power supply is highly desirable to both customers and utility companies. Since faults in distribution systems account for the majority of customer reliability problems [1], many studies have been devoted to distribution faults.

To locate the fault sections, relay alarms, SCADA measurements, Automated Meter Reading (AMR) messages and customer calls are used [3]. More accurate fault locations can be estimated from real-time fault current and voltage waveforms by extracting their frequency features [4] or matching with the values calculated from system models [5]. Due to

safety concerns, the affected power lines can only be reenergized after the root cause of the fault is identified and cleared. This is mostly done by distribution engineers. They have to go on-site, drive or walk miles along the power line, collect evidences for the root cause.

To expedite this process, automated fault diagnosis has been studied to provide engineers with information about the root cause. Automated fault diagnosis learns the relationship between fault properties and its root cause using historical fault events. For example, Chen *et al*. [6] proposed an online diagnosis approach using cause-effect network and fuzzy rules to find out the root cause based on protective device settings and their operations during the fault in distribution substations; Zhang *et al*. [7] and Hor *et al*. [8] proposed to use Rough Set theory to discover and represent the relationship between faults and the sequence of protective actions; Guikema *et al*. [9] and Bernardi *et al*. [10] built statistical models to relate fault to tree trimming and lightning; Gui, Pahwa and Das [11] analyzed the trend and annual pattern of animal-caused faults in Kansas; Nunez *et al*. [12] used features extracted from the fault current and voltage waveforms and environmental factors to diagnose the root cause of distribution faults on overhead lines; Xu and Chow *et al*. [13-15] formulated fault diagnosis as a classification problem and applied several biologically inspired algorithms, including artificial neural networks (ANN), fuzzy systems and artificial immune recognition systems (AIRS).

II. SMALL WORLD STRATIFICATION

The origin of small world concept is an observation in examining the average path length for social network of people in the United States. Several experiments in 1960's [16, 17]

suggest that human society is a small-world network with a very short path length – we can find a person through on average six mutual friends, which is frequently referred to as "six degrees of separation".

A small-world network is characterized by a small average shortest path length and a large clustering coefficient [18]. In other words, a small-world network is a network where nodes are relatively close to each other in the sense that only a small number of hops are needed to reach each other although the connections are globally sparse, and where local clusters formed by densely connected nodes exist.

Many real-world networks, such as the Internet, World Wide Web, and Criminal Activity Network, show small-world properties in their topology [19, 20]. The small average shortest path length of small-world networks is applied in the design of fast routing and searching schemes in peer-to-peer networks [21-23]. In power engineering, the small-world networks have been explored in various perspectives as well. The early work mainly focused on characterizing the topology of power grids using the small-world network model [18, 24]. The recent research by Wang, Scaglione and Thomas applied the small-world network model to constrain the topology of generated power lines for simulation purposes [25]. The vulnerability of power grids, i.e. whether the grid can sustain safe operation with failure of nodes, was investigated from the network connection perspective as well [26].

Small World Stratification (SWS) discussed in this dissertation concerns the sampling issues in distribution fault diagnosis. As distribution systems are spatially dispersed and heterogeneous, instead of digging into the entire outage database, historical fault records are sampled to investigate a certain fault event most of time. The sampling involves a time

window, such as the past a couple of years or certain months prior to the fault event. Meanwhile, the sampling limits the fault events within a spatial extent, which could be as small as the service region of a distribution lateral or as big as the distribution systems of a major city and neighboring towns.

The sampling process affects the study of fault-environment relationships significantly. Diagnosis of the root cause of a fault event needs recent information about the local environments. Therefore, fault events that are too old or too far away from the system under investigation generally provide little useful information. In different local systems, even the environmental features needed to diagnose the fault root cause could be different. Thus, proper sampling of historical fault events is crucial in automated fault diagnosis.

Intuitively, investigating fault events within a small geographic area is preferable to focus on the local fault characteristics. However, a small study area may lead to too few historical fault events for the fault diagnosis algorithms to learn the relationship between fault root causes and the environments. SWS tackles this problem by sampling additional fault event from other areas while keeping the study area small.

Outage records from the distribution systems of Progress Energy Carolinas Inc. are first integrated with environmental data from other resources and then analyzed to reveal the effect of sampling. A power distribution fault simulator is developed to generate realistic fault events for various experiments in this dissertation. The characteristics of the simulated fault events are validated with the real-world outage records and the proposed SWS algorithm is tested on the actual outage records as well.

III. OVERVIEW OF THE DISSERTATION

This dissertation consists of several manuscripts submitted for publication to journals and conferences. Chapter II is published in the proceedings of the 2009 IEEE Power and Energy Society General Meeting [27]. Chapter III is published in IEEE Transactions on Power Systems [28]. Chapter IV is accepted for publication in IEEE Transactions on Power Systems [29]. Chapter V through Chapter VII has been submitted to IEEE Transactions on Power Systems and is currently under review. There are also several other conference papers that have been published or accepted for presentation along the research conducted in this dissertation [30-33].

Chapter II introduces using Geographic Information System (GIS) as a framework to integrate data from various sources through spatial and temporal relations. Tools for exploratory data analysis, including likelihood measure, Linear Discriminant Analysis (LDA), and Logistic Regression (LR), are introduced and compared.

Chapter III focuses on selecting significant features for fault diagnosis from massive data. This chapter reviews two popular feature selection methods: a) hypothesis test, b) stepwise regression, and introduces another two: c) stepwise selection by Akaike's Information Criterion, and d) LASSO/ALASSO. These four methods are compared in terms of their model requirements, data assumptions, computational cost, and fault diagnosis performance on real-world datasets.

To facilitate the investigation on sampling strategies, Chapter IV proposes a framework for modeling and simulating fault events in power distribution systems based on environmental factors and the cause-effect relationships among them. The spatial and

temporal aspects of significant environmental factors leading to various faults are modeled as raster maps and probability functions, respectively. The cause-effect relationships are expressed as fuzzy rules and a hierarchical fuzzy inference system is built to infer the probability of faults in the simulated environments.

Chapter V through Chapter VII proposes the Small World Stratification (SWS) sampling strategy and details the algorithms design and implementation. Chapter V explains the concept of SWS and uses fault events simulated by the Distribution Fault Simulator [29] to demonstrate its effectiveness. Chapter VI proposes four regional feature vectors (RFV) derived from measures used to analyze distribution faults and evaluates similarity measures based on the distance between RFVs. Chapter VII details the algorithm design and implementation of SWS.

REFERENCES

[1] R. E. Brown, Electric *Power Distribution Reliability*. New York: Marcel Dekker, Inc, 2002.

[2] Electricity Advisory Committee, US Department of Energy "Smart grid: enabler of the new energy economy," [Online]. Available: http://www.oe.energy.gov/DocumentsandMedia/final-smart-grid-report.pdf.

[3] K. Sridharan and N. Schulz, "Outage management through AMR systems using an intelligent data filter," *IEEE Power Engineering Review*, vol. 21, p. 64, 2001.

[4] 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 Systems*, vol. 23, pp. 380-388, 2008.

[5] S.-J. Lee, M.-S. Choi, S.-H. Kang, B.-G. Jin, D.-S. Lee, B.-S. Ahn, N.-S. Yoon, H.-Y. Kim, and S.-B. Wee, "An intelligent and efficient fault location and diagnosis scheme for radial distribution systems," *IEEE Trans. Power Delivery*, vol. 19, pp. 524-532, 2004.

[6] W.-H. Chen, C.-W. Liu, and M.-S. Tsai, "On-line fault diagnosis of distribution substations using hybrid cause-effect network and fuzzy rule-based method," *IEEE Trans. Power Delivery*, vol. 15, pp. 710-717, 2000.

[7] Q. Zhang, Z. Han, and F. Wen, “A new approach for fault diagnosis in power systems based on rough set theory,” presented at the 4th Int. Conf. Advances in Power System Control, Operation and Management, Hong Kong, China, 1997.

[8] C.-L. Hor, P. A. Crossley, and S. J. Watson, "Building knowledge for substation-based decision support using rough sets," *IEEE Trans. Power Delivery*, vol. 22, pp. 1372-1379, 2007.

[9] S. D. Guikema, R. A. Davidson, and H. Liu, "Statistical models of the effects of tree trimming on power system outages," *IEEE Trans. Power Delivery*, vol. 21, pp. 1549-1557, 2006.

[10] M. Bernardi, A. Borghetti, C. A. Nucci, and M. Paolone, "A statistical approach for estimating the correlation between lightning and faults in power distribution systems," in *2006 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pp. 1-7.

[11] M. Gui, A. Pahwa, and S. Das, "Analysis of animal-related outages in overhead distribution systems with wavelet decomposition and immune systems-based neural networks," *IEEE Trans. Power Systems*, vol. 24, pp. 1765-1771, 2009.

[12] V. B. Nunez, S. Kulkarni, S. Santoso, and J. Melendez, "Feature analysis and classification methodology for overhead distribution fault events," in *IEEE Power and Energy Society General Meeting 2010*, Minneapolis, MN, 2010.

[13] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Trans. Power Systems*, vol. 21, pp. 53-60, 2006.

[14] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm," *IEEE Trans. Power Systems*, vol. 22, pp. 164-171, 2007.

[15] L. Xu, M.-Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using Artificial Immune Recognition System (AIRS) algorithm," *IEEE Trans. Power Systems*, vol. 22, pp. 198-204, 2007.

[16] S. Milgram, "The small world problem," *Psychology Today*, vol. 2, pp. 60-67, 1967.

[17] J. Travers and S. Milgram, "An experimental study of the small world problem," *Sociometry*, vol. 32, pp. 425-443, 1969.

[18] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-442, 1998.

[19] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167-256, 2003.

[20] S. Kaza, J. Xu, B. Marshall, and C. Hsinchun, "Topological analysis of criminal activity networks: enhancing transportation security," *IEEE Trans. Intelligent Transportation Systems*, vol. 10, pp. 83-91, 2009.

[21] G. Chen, C. P. Low, and Z. Yang, "Enhancing search performance in unstructured P2P networks based on users' common interest," *IEEE Trans. Parallel and Distributed Systems*, vol. 19, pp. 821-836, 2008.

[22] M. Li, W.-C. Lee, A. Sivasubramaniam, and J. Zhao, "SSW: a small-world-based overlay for peer-to-peer search," *IEEE Trans. Parallel and Distributed Systems*, vol. 19, pp. 735-749, 2008.

[23] H.-C. Hsiao, Y.-C. Lin, and H. Liao, "Building small-world peer-to-peer networks based on hierarchical Structures," *IEEE Trans. Parallel and Distributed Systems*, vol. 20, pp. 1023-1037, 2009.

[24] P. Crucitti, V. Latora, and M. Marchiori, "A topological analysis of the Italian electric power grid," *Physica A*, vol. 388, pp. 92-97, 2004.

[25] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating statistically correct random topologies for testing smart grid communication and control networks," *IEEE Trans. Smart Grid*, to be published.

[26] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the North American power grid," *Physical Review E*, vol. 69, pp. 1-4, 2004.

[27] Y. Cai and M.-Y. Chow, "Exploratory analysis of massive data for distribution fault diagnosis in Smart Grids," presented in *IEEE Power & Energy Society General Meeting,* Calgary, Canada, 2009.

[28] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Statistical feature selection from massive data in distribution fault diagnosis," *IEEE Trans. Power Systems*, vol. 25, pp. 642-648, 2010.

[29] Y. Cai and M.-Y. Chow, "Cause-effect modeling and spatial-temporal simulation of power distribution fault events," *IEEE Trans. Power Systems*, to be published.

[30] Y. Cai and M.-Y. Chow, "Cause-effect modeling and simulation of power distribution fault events," in *Proc. IEEE Power & Energy Soc. General Meeting*, Minneapolis, MN, 2010.

[31] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Evaluation of distribution fault diagnosis algorithms using ROC curves," in *Proc. IEEE Power & Energy Soc. General Meeting*, Minneapolis, MN, 2010.

[32] Y. Cai and M.-Y. Chow, "Small world stratification for distribution fault diagnosis," accepted to present at *2011 Power Systems Conference & Exposition*, Phoenix, AZ.

[33] Y. Cai and M.-Y. Chow, "Similarity measures in small world stratification for distribution fault diagnosis," accepted to present at *2011 Power Systems Conference & Exposition*, Phoenix, AZ.

CHAPTER II

EXPLORATORY ANALYSIS OF MASSIVE DATA FOR DISTRIBUTION FAULT DIAGNOSIS IN SMART GRIDS

Yixin Cai Mo-Yuen Chow

ycai2@ncsu.edu chow@ncsu.edu

Department of Electrical and Computer Engineering

North Carolina State University

Raleigh, NC 27695 USA

This chapter is accepted for presentation at 2009 IEEE Power and Energy Society General Meeting (PESGM '09)

EXPLORATORY ANALYSIS OF MASSIVE DATA FOR DISTRIBUTION FAULT DIAGNOSIS IN SMART GRIDS

Abstract -- Fault diagnosis in power distribution systems is critical to expedite the restoration of service and improve the reliability. With power grids becoming smarter, more and more data beyond utility outage database are available for fault cause identification. This paper introduces basic methodologies to integrate and analyze data from different sources. Geographic Information System (GIS) provides a framework to integrate these data through spatial and temporal relations. Features extracted from raw data provide different discriminant powers, which can be evaluated by the likelihood measure. A fault cause classifier is then trained to learn the relations between fault causes and the features. Two statistical methods, Linear Discriminant Analysis (LDA) and Logistic Regression (LR), are introduced. The assumptions, general approaches and performances of these two techniques are discussed and evaluated on a real-world outage dataset.

Index Terms — Classification, Fault Cause Identification, Geographic Information System, Power Distribution Systems, Smart Grids, Spatial-Temporal Relation

I. INTRODUCTION

As the retail part of power systems, distribution systems are characterized by large scale multi-branched radial topologies, geographically dispersed components, and non-linear

operation dynamics. Due to exposure to harsh and uncertain environment, outages in distribution systems account for up to 90% of all customer reliability problems [1]. In order to minimize the fault-induced losses for customers and improve the system reliability, it is important to locate the fault, identify the causes and restore the service in a timely manner.

Outage Management Systems (OMS) have been deployed in utilities for years to facilitate post-fault reactions. Relay alarms, SCADA measurements, Automated Meter Reading (AMR) messages and customer calls are used to inference fault sections [2]. More accurate fault locations can be estimated from real-time fault current and voltage waveforms by extracting their frequency features [3] or matching with the values calculated from system models [4]. However, fault cause identification has been heavily relied on human experience and intelligence. Field engineers have to go on-site, drive or walk miles along the power line, find clues to locate the fault spot, and collect evidences for the root cause. As the service territories vary, root causes of concern are different in different utilities. Equipment failures, trees, lightning strikes and animals are considered to be the primary causes of faults in distribution systems [1]. The complex and stochastic nature of environmental features makes it very difficult to identify the cause only by electrical measurements, such as currents. Information beyond the circuit is needed.

Various efforts have been made to establish relationship between environmental features and fault causes. For example, Guikema *et al*. [5] and Bernardi *et al*. [6] built statistical models to relate fault to tree trimming and lightning; Xu *et al*. [7] and Sahai *et al*. [8] characterized and modeled features related to tree and animal faults by logistic regression and simple Bayesian networks. Moreover, automated fault cause identification has also been

investigated. Chen *et al*. [9] proposed an online diagnosis approach using cause-effect network and fuzzy rules to find out the root cause based on protective device settings and their operations during the fault in distribution substations; Xu, Chow *et al*. [10-12] formulated distribution fault cause identification as a classification problem and applied various biologically inspired algorithms, such as neural network, fuzzy system and Artificial Immune Recognition Systems (AIRS).

Smart grids are the vision of future power systems, where the applications of communication and information technologies enable more reliable, efficient and environment friendly deliver of electric power [13]. The deployment of advanced sensors and communication technologies in smart grids would provide richer, more accurate, and real-time data about the system states and its environments, which will change the way of fault diagnosis and eventually enable automated fault cause identification.

This paper will discuss methodologies for integrating and analyzing new data obtained beyond conventional OMS, which will be effective technologies for fault management in smart grid environments. The methods are illustrated with examples from a real-world outage dataset. The rest of the paper is structured as follows: Section II introduces common data sources and the way to integrate data by spatial-temporal relations under GIS framework; Section III discusses how to analyze a single feature and choose the influential one; Section IV introduces building a simple fault cause classifier with Linear Discriminant Analysis and Logistic Regression; Section V is the conclusion.

II. DATA INTEGRATION

A. Common Data Sources

The amount of data related to distribution faults are fast growing due to the rapid development of information technology. The conventional and most essential data is historical fault records. Most utilities keep recording outages in their systems and have accumulated large databases over years. In such a database, one outage is associated with tens of fields concerning its properties. Common information, such as time of the fault occurrence, circuit and protective devices involved, fault duration, and customers affected, is recorded. The root causes, after being identified by field engineers, is also recorded with detailed explanations. This database provides the basis of how different faults behave.

Nowadays, various environmental data are available online. The major sources are government agencies. To name a few: historical and real-time weather measurements (e.g. temperature, wind speed, precipitation) from more than 3000 weather sites in North Carolina are available from NC CRONOS database maintained by North Carolina State Climate Office; the land use and land cover with a spatial resolution of 30 meters for the entire US can be retrieved from National Land Cover Database jointly supported by tens of federal agencies; most local governments provide data service on major geographic features, such as roads, water bodies and aerial photos. Data with specific themes is available from commercial vendors as well.

B. GIS as a Data Integration Framework

Geographic Information System (GIS) is an information system capable of integrating, storing, editing, analyzing, sharing, and displaying geographically referenced information

[14]. It has been used in power engineering for asset management, system visualization and system planning. It is also an integrated part of many Distribution Management Systems (DMS) and OMS to maintain digital maps and dispatch repair crews. However, the capability of GIS as an analysis tool in distribution fault cause identification is yet to be explored.

A fault occurs somewhere at some time, so location and time are two fundamental properties of a fault event. Kezunovic and Abur [15] proposed using spatial-temporal relation to integrate SCADA and IED data to improve power system monitoring. More generally, a mathematical framework, space-time point process, is used to describe the events occurring at random locations in space and random time instant. The spatial-temporal relations behind certain events such as earthquakes [16] and city crimes [17] are investigated to characterize the event and predict future occurrences. This type of relations is also applicable in distribution fault diagnosis. Data from other sources can be integrated into utility outage databases in the following ways:

- Spatial relation – new data collected around where the fault occurred is appended to the outage record. This type of data is usually considered to be static, which changes slowly with time or does not change at all within the study time period. In a GIS system, this relation could be a spatial join of a point (fault location) to the nearest point (e.g. weather stations), polyline (e.g. highways) or polygon (e.g. vegetation, land use). Fig. 1 is an example of integrating data by spatial relation, where a land use type is assigned to a fault by the land use where it occurred.
- Spatial-temporal relation – new data is restricted by the time when the fault occurred besides spatial proximity. This type of dynamic data changes over time and need to

be sampled properly. Weather measurements belong to this category and a time stamp match is required to link the new data to the outage record. The matching operation can be implemented through joining tables in a database.

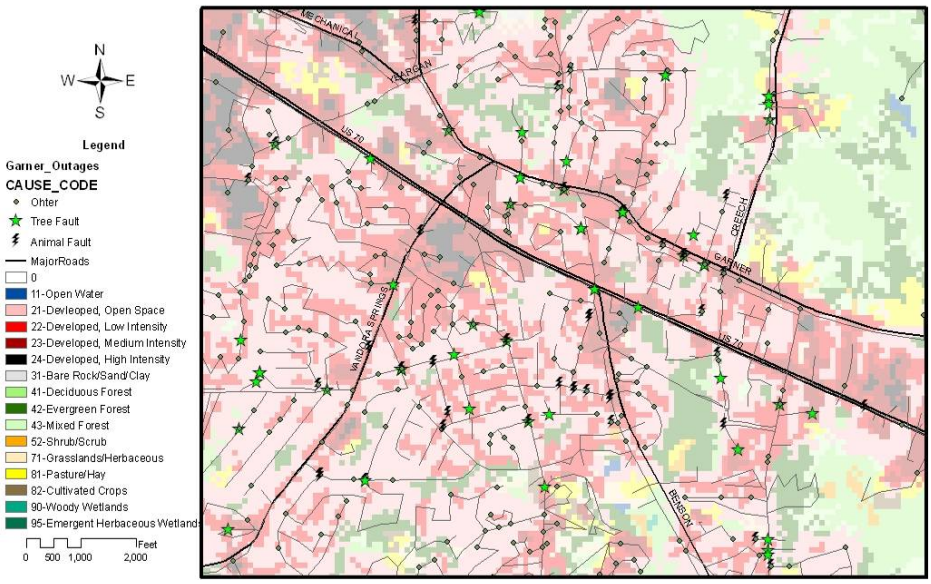


Fig. 1. Land use type is assigned according to where the fault occurred.

GIS provides a good framework and a powerful software tool to handle both the time and spatial relations. For example, the current version of ArcGIS by ESRI supports most of common database formats. Both ordinary database queries and geographical queries and processing can be performed under the same software using one single database.

III. Evaluate A Single Feature

A. Data Preprocess

Raw data collected from utilities and other sources are often noisy and may contain errors due to device malfunctions or human mistakes that are needed to be cleaned before analysis. The following processing is usually performed:

1. Define the root causes of interest.

For various analysis and management purposes, utilities have defined detailed outage causes. For example, 50 outage causes in 8 categories are specified in Progress Energy Carolinas outage dataset. However, many of these outages are not caused by faults, such as planned outage for system maintenance and upgrade. While some of them are out of our research scope, such as outages caused by customer side equipment and transmission level failures. These outages are not considered in our analysis.

Outages caused by faults need to be regrouped based on different focuses. For example, if we are interested in tree faults, all the vegetation related faults, including limb contact, fallen branches, etc. are grouped as tree faults, and all the other faults are grouped as one class, non-tree faults.

2. Clean errors and noises.

Some obvious errors in raw data can be filtered out. For categorical data, values that are not defined should be deleted. For continuous data, meaningless values such as negative wind speed should be filtered out.

Most noises are difficult to be filtered because usually there is little information about how the data was collected. Under certain conditions, some outliers in continuous measurements can be detected. For instance, a jumping value in consecutive hourly temperature readings could be identified as an outlier.

3. Extract features.

Not all the fields in an outage database are useful or appropriate for fault diagnosis or analysis. So selecting data fields and converting them to features are necessary.

Categorical data, such as protective device type, can often be directly used as features. Reclassification of categories is sometimes necessary following preliminary analysis. Many other data need to be converted to another form for information extraction. For instant, the date when a fault occurred could be converted to season in order to reflect the yearly cyclic pattern; measured wind speed could be converted to hourly average and maximum value to reduce the data amount.

B. Analyze Categorical Features

A categorical feature takes only a limited number of discrete values. It can be nominal, which means the values are not ordered. It can be ordinal, which means the observed levels are ordered in some way.

A good method to analyze categorical features is the likelihood measure [18]. It is essentially the conditional probability of certain type of faults, given a specific value of a feature. The likelihood measure is defined as

$$L_{i,j} = P(o_i \mid X = x_j) = \frac{N_{i,j}}{N_j}, \tag{1}$$

where $N_{i,j}$ represents the number of type i faults and N_j represents the number of all faults, given feature value $X = x_j$.

A visual tool for categorical feature analysis is Mosaic plot as shown in Fig. 2. The x-axis represents feature values and y-axis gives the percentage of samples in different classes. The block area is proportional to the number of samples. Take Mosaic plot of tree faults against weather conditions as an example. We can see that most faults occurred under weather conditions 1 (clear weather) while very few are under weather condition 2 (extreme

temperature). The likelihood of tree faults can be read directly from the height of the bottom bars (y equals to 1). In this example, tree faults are most likely under weather condition 8 (windy) with a value up to 0.7 and least likely under weather condition 1 (clear weather).

Fig. 2. Mosaic plots for categorical features.

Compared to weather conditions, Mosaic plot of tree faults in different seasons reveals quite different characteristics: numbers of tree faults in four seasons are similar while likelihood of tree faults is only a little higher in season 2 (summer) than others.

If we have to make a decision on the fault cause based on one single feature, different discriminant power will appear. With weather conditions, we would classify a fault as a tree fault under weather condition 8 and non-tree fault otherwise. With seasons, we could not make any decision better than taking them all as non-tree fault. Intuitively, a feature with significantly different likelihood under different values provides better discrimination power in fault cause identification. Thus, as shown in Fig. 2, discriminant power of land use data is better than season but not as good as weather conditions.

C. Analyze Continuous Features

Likelihood measure for continuous features is a little complicated. It could be approximated by dividing the raw data value into several bins and calculate the conditional probabilities in each bin. However, the result depends on the size of bins and does not have a clear meaning. Therefore, we define the likelihood measure of a continuous feature as

$$L_{i,j} = P(o_i \mid X \geq x_j) = \frac{N_{i,j}}{N_j}, \tag{2}$$

where notations are with the same meaning as in (1).

This definition is also a conditional probability. The calculation only depends on the feature value without any other parameters.

Fig. 3 gives examples of likelihood measure of continuous features. The starting point at the left end indicates approximately the overall likelihood of this type of faults. Trends of likelihood as the feature value increases can be observed before the feature value goes too large. With a large feature value, usually there is not enough fault cases to support a reliable likelihood measure. This explains the sparse points and fluctuation at the right end.

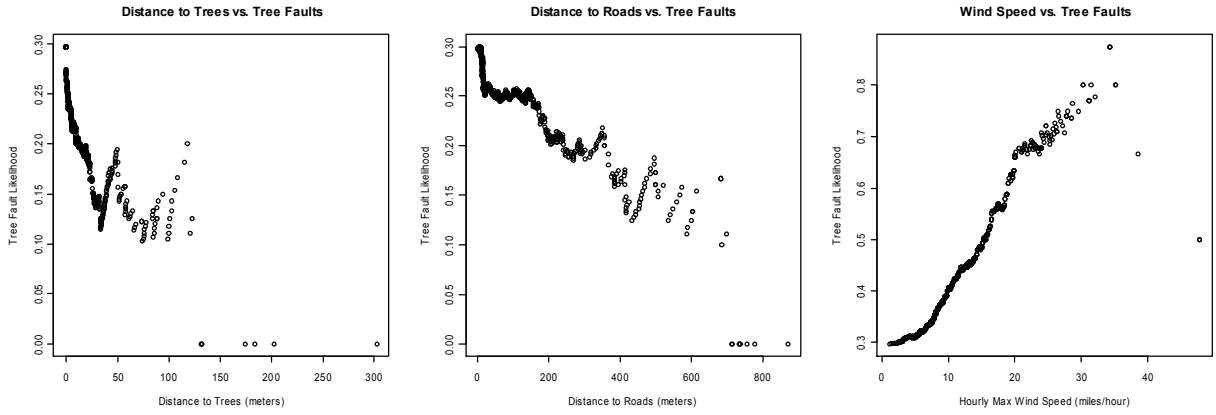


Fig. 3. Likelihood measure of continuous features.

As shown in Fig. 3, the likelihood of tree faults decreases as it is farther away from trees and roads, but increases rapidly as the wind speed increases. Note that the range of likelihood change is quite different. Similar to categorical features, a feature with significantly different likelihood under different values is preferred for fault cause identification.

Wind speed provides stronger discriminant power than distance to trees and road in this example. Thus, it is a feature that should carry more weight in fault cause identification process.

IV. BUILD A FAULT CAUSE CLASSIFIER

Distribution fault cause identification is formulated as a classification problem in our study. Each fault event can be represented as a vector $[c, f_1, \cdots, f_N]^T$. The classification variable c takes nominal values indicating the root causes of interest, such as tree and animal. The quantitative variables $\mathbf{f} = [f_1, \cdots, f_N]^T$ are features associated with this event. Classification algorithms are used to decide the value of c given the values of $\mathbf{f}$. In this paper, we will focus on the discussion of the two-class (binary) problem.

A. Linear Discriminant Analysis

Discriminant Analysis (LDA) is a common method for classification [19]. It builds a linear combination of the features based on training set – a set of samples with known classes and feature values. This combination, usually referred to as discriminant function

$$D = \mathbf{w}^T \mathbf{f} = \sum_{i=1}^{N} w_i f_i \,, \tag{3}$$

is used for classifying new samples whose class is unknown. In a two-class case, the sample is assigned to one class when the discriminant function value exceeds a certain threshold.

LDA assumes that **f** is a multivariate normal distribution, and the within class variance $\sum_c$ is equal for different classes. The Bayesian optimal discriminant function for a two-class problem can be derived as

$$D = [\sum^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)]^T \mathbf{f}\,, \tag{4}$$

where $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ are within-class mean of **f**.

B. Logistic Regression

Logistic Regression (LR) is another well known statistic method to analyze problems with binary dependent variable [20]. LR builds a model to predict the logarithm odds for the sample to be in one class.

Suppose the two classes under study are $c \in \{0,1\}$, the logistic regression model is

$$\text{logit}(c=1) = \ln\frac{P(c=1)}{P(c=0)} = \alpha + \boldsymbol{\beta}^T \mathbf{f}\,, \tag{5}$$

where α and $\boldsymbol{\beta}$ are unknown parameters to be identified with maximum likelihood method using the training set.

Equation (5) can be easily solved for the probability of the observed sample being in the class of interest:

$$P(c=1) = \frac{1}{1+e^{-\alpha-\boldsymbol{\beta}^T\mathbf{f}}}\,. \tag{6}$$

The class is assigned by comparing the calculated probability with the predefined threshold, naturally 0.5 in two-class problem. An optimal threshold to minimize the classification error can be found through experimentations.

C. Comparison of LDA and LR

LDA and LR fit models with training data to predict the class of an observed sample. From the application point of view, they are similar to each other with the following differences:

- Model -- LDA finds linear combination of features and is essentially a linear classifier which represents a hyper plain in the feature space. LR fits a linear model for logarithm odd so is non-linear in term of classifiers.
- Data assumption – LDA assumes normality and equal variance of features, which are very strong assumptions and can be easily violated by real-world data. In contrary, LR takes both categorical and continuous features as regressors and does not require the distribution or variance.
- Computational cost -- LR uses maximum likelihood method to estimate the coefficients. With a fairly large dataset, this iterative algorithm requires more computation than matrix manipulations used in LDA.

In general, LR is more favorable in our case because categorical features are widely found in outage database. The actual classification performance of LDA and LR will be compared on a real-world outage dataset.

D. Case Study

The dataset used in this case study is from the outage database of Progress Energy Carolinas. Outages between 2005 and 2006 in an area around the city of Raleigh are studied (as shown in Fig.4). Tree and animal faults are our focuses.

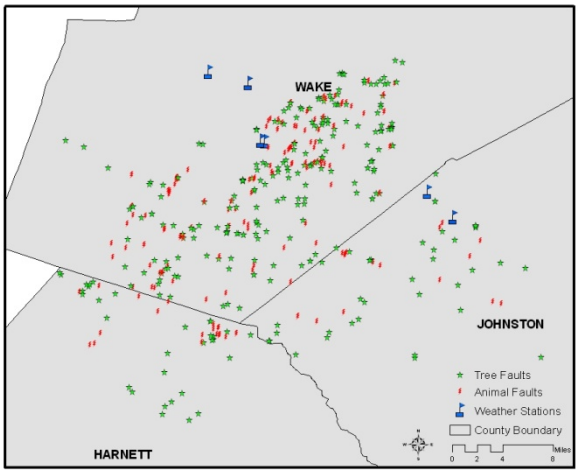


Fig. 4. Study region and faults.

Six categorical features are extracted from the outage database: number of phases affected, season, time of the day, protective device type, weather conditions and an indicator of overhead or underground circuit. Another 6 features are obtained from other data sources. Land use, distance to trees and distance to roads are integrated by spatial relation. Daily average temperature, daily precipitation and hourly maximum wind speed from the nearest weather stations are integrated by spatial-temporal relation. Except land use, they are all continuous features.

Due to outage data imbalance, using classification accuracy (ACC) as the only performance measure could be misleading [11]. Therefore, two more performance measures are used. Suppose the class we are interested in (e.g. tree fault) is positive class, probability

of detection (POD) is a measure of how well the positive class is correctly found and false alarm ratio (FAR) represents how well we can believe it when a sample is classified as positive class [21]. In a perfect case, ACC and POD would be 1, and FAR equals to 0.

Table I and II list the experimental results with LR and LDA on the sample dataset. Classification using 6 features from the outage database is compared to the 11 features including those obtained from other data sources. Note that categorical weather conditions are not included in the 11 features in order to avoid collinearity. The sample dataset is randomly split as training set and testing set with equal size. Models are built based on training set and used to classify both. Experiments are repeated 30 times and the average performance measures are reported in the table with standard deviation in brackets.

TABLE I

CLASSIFICATION PERFORMANCE USING LDA ON SAMPLE DATASET

		6 Features		11 Features	
		training	testing	training	testing
Tree fault	ACC	0.75(0.01)	0.76(0.01)	0.77(0.02)	0.76(0.02)
	POD	0.32(0.03)	0.34(0.03)	0.41(0.03)	0.39(0.03)
	FAR	0.34(0.03)	0.32(0.03)	0.32(0.04)	0.33(0.04)
Animal fault	ACC	0.84(0.02)	0.83(0.01)	0.84(0.02)	0.84(0.01)
	POD	0.31(0.04)	0.29(0.04)	0.35(0.03)	0.35(0.03)
	FAR	0.42(0.05)	0.43(0.05)	0.39(0.05)	0.41(0.05)

TABLE II

CLASSIFICATION PERFORMANCE USING LR ON SAMPLE DATASET

		6 Features		11 Features	
		training	testing	training	testing
Tree fault	ACC	0.76(0.02)	0.76 (0.02)	0.77 (0.01)	0.77(0.02)
	POD	0.32(0.03)	0.32(0.03)	0.44 (0.03)	0.44(0.03)
	FAR	0.30(0.04)	0.30(0.04)	0.32 (0.03)	0.34(0.03)
Animal fault	ACC	0.83(0.02)	0.83 (0.02)	0.84(0.01)	0.84(0.01)
	POD	0.30(0.03)	0.31(0.03)	0.37 (0.04)	0.35(0.03)
	FAR	0.42(0.04)	0.41(0.06)	0.41 (0.06)	0.41(0.06)

From the tables we can observe:

- The performance on testing set is very close to the training set. This means both LDA and LR provide good predictability, and there is no over-fitting problem.
- The performance difference between LDA and LR is negligible. Although categorical features are used in LDA, this violation of assumption does not affect the classification much.
- With new features added, the classification accuracy does not change. But POD is improved while maintaining FAR at the same level. These new features are helpful to fault cause identification.

V. CONCLUSION

With power grids becoming smarter, more and more data is available for fault cause identification. This paper introduces basic methodologies to integrate and analyze data from different sources.

Historical fault records are usually available in utilities and a huge amount of additional data can be retrieved online. GIS provides a framework to integrate these data through spatial and temporal relations.

The first step in data analysis is cleaning up the raw data by defining root causes of interest, filtering out errors and extracting features. The discriminant power provided by a single feature can be characterized by the likelihood measure. A feature with significantly different likelihood under different values is preferred in fault cause identification.

The next step is to build a classifier which can learn fault characteristics from historical records. Linear Discriminant Analysis and Logistic Regression are two widely used methods for classification. Although fewer assumptions are required for Logistic Regression, both methods achieved similar performance in our experiments with mixed categorical and continuous features.

Adding appropriate features from other data sources can improve classification performance in terms of detection power. However, the overall performance might still have low POD and high FAR. To further improve the fault cause identification performance, more systematic feature selection methods, new classifiers as well as novel sampling strategies to cope with data imbalance are needed.

VI. ACKNOWLEDGMENT

The authors gratefully acknowledge the contribution of John W. Gajda and Glenn C. Lampley from Progress Energy Carolinas Inc. for their support on data and field experience.

REFERENCES

[1] R. E. Brown, *Electric Power Distribution Reliability*. New York: Marcel Dekker, Inc, 2002.

[2] K. Sridharan and N. Schulz, "Outage management through AMR systems using an intelligent data filter," , IEEE *Power Engineering Review*, vol. 21, p. 64, 2001.

[3] 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 Systems*, vol. 23, pp. 380-388, 2008.

[4] S.-J. Lee, M.-S. Choi, S.-H. Kang, B.-G. Jin, D.-S. Lee, B.-S. Ahn, N.-S. Yoon, H.-Y. Kim, and S.-B. Wee, "An intelligent and efficient fault location and diagnosis scheme for radial distribution systems," *IEEE Trans. Power Delivery*, vol. 19, pp. 524-532, 2004.

[5] S. D. Guikema, R. A. Davidson, and H. Liu, "Statistical models of the effects of tree trimming on power system outages," *IEEE Trans. Power Delivery*, vol. 21, pp. 1549-1557, 2006.

[6] M. Bernardi, A. Borghetti, C. A. Nucci, and M. Paolone, "A statistical approach for estimating the correlation between lightning and faults in power distribution systems," in *2006 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pp. 1-7.

[7] L. Xu, M.-Y. Chow, and L. S. Taylor, "Analysis of tree-caused faults in power distribution systems," presented in the 35th North American Power Symposium, University of Missouri-Rolla in Rolla, Missouri, 2003.

[8] S. Sahai and A. Pahwa, "A probabilistic approach for animal-caused outages in overhead distribution systems," in *2006 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS),* pp. 1-7.

[9] W.-H. Chen, C.-W. Liu, and M.-S. Tsai, "On-line fault diagnosis of distribution substations using hybrid cause-effect network and fuzzy rule-based method," *IEEE Trans. Power Delivery*, vol. 15, pp. 710-717, 2000.

[10] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Trans. Power Systems*, vol. 21, pp. 53-60, 2006.

[11] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-algorithm," *IEEE Trans. Power Systems*, vol. 22, pp. 164-171, 2007.

[12] L. Xu, M.-Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using artificial immune recognition system (AIRS) algorithm," *IEEE Trans. Power Systems*, vol. 22, pp. 198-204, 2007.

[13] US Department of Energy, "The smart grid: an introduction," [Online]. Available: http://www.oe.energy.gov/DocumentsandMedia/ DOE_SG_Book_Single_Pages.pdf

[14] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, *Geographic Information Systems and Science*. Chichester, England: John Wiley & Sons, Ltd., 2001.

[15] M. Kezunovic and A. Abur, "Merging the temporal and spatial aspects of data and information for improved power system monitoring applications," *Proceedings of the IEEE*, vol. 93, pp. 1909-1919, 2005.

[16] Y. Ogata, "Space-time point process models for earthquake occurrences," *Annals of the Institute of Statistical Matheatics*, vol. 50, pp. 379-402, 2004.

[17] H. Liu and D. E. Brown, "A new ponit process transition densigy model for space-time event prediction," *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 34, 2004.

[18] M.-Y. Chow and L. S. Taylor, "A novel approach for distribution fault analysis," *IEEE Trans. Power Delivery*, vol. 8, pp. 1882-1889, 1993.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed: John Wiley & Sons, 2001.

[20] R. L. Ott and M. Longnecker, *An Introduction to Statistical Methods and Data Analysis*, 5th ed. Pacific Grove, CA, USA: Duxbury, 2001.

[21] WWRP/WGNE Joint Working Group on Verification, "Forecast verification - issues, methods and FAQ," [Online]. Available: http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.htm.

CHAPTER III

STATISTICAL FEATURE SELECTION FROM MASSIVE DATA IN DISTRIBUTION FAULT DIAGNOSIS

Yixin Cai[1]	Mo-Yuen Chow[1]	Wenbin Lu[2]	Lexin Li[2]
ycai2@ncsu.edu	chow@ncsu.edu	wlu4@stat.ncsu.edu	li@stat.ncsu.edu

1. Department of Electrical and Computer Engineering

2. Department of Statistics

North Carolina State University

Raleigh, NC 27695 USA

This chapter is published in IEEE Transactions on Power Systems,

vol. 25, no. 2, pp. 642-648, 2010

STATISTICAL FEATURE SELECTION FROM MASSIVE DATA IN DISTRIBUTION FAULT DIAGNOSIS

Abstract — Selecting proper features to identify the root cause is a critical step in distribution fault diagnosis. Power engineers usually select features based on experience. However, engineers cannot be familiar with every local system, especially in fast growing regions. With the advancing information technologies and more powerful sensors, utilities can collect much more data on their systems than before. The phenomenon will be even more substantial for the anticipating Smart Grid environments. To help power engineers select features based on the massive data collected, this paper reviews two popular feature selection methods: a) hypothesis test, b) stepwise regression, and introduces another two: c) stepwise selection by Akaike's Information Criterion, and d) LASSO/ALASSO. These four methods are compared in terms of their model requirements, data assumptions, and computational cost. With real-world datasets from Progress Energy Carolinas, this paper also evaluates these methods and compares fault diagnosis performance by accuracy, probability of detection and false alarm ratio. This paper discusses the advantages and limitations of each method for distribution fault diagnosis as well.

Index Terms -- Akaike's information criteria, classification, fault cause identification, feature selection, hypothesis test, power distribution systems, LASSO, logistic regression, smart grid, stepwise regression

I. INTRODUCTION

Power systems are vital lifelines of the modern society for maintaining adequate and reliable flows of energy. To cover their service territory, power lines extend thousands of miles, link numerous voltage conversion equipment and protective devices together, and form an interconnected network. As the retail part of power systems, typical distribution systems are geographically dispersed and hence exposed to harsh and uncertain environments. Therefore, the systems can easily be affected by various outage-causing events such as equipment failures, animal contacts, trees, lightning strikes, etc [1]. It is important to diagnose the faults and restore the service in a timely manner in order to minimize the fault-induced losses. However, due to the complex and stochastic nature of environmental factors, fault cause identification has been heavily relied on human experience and intelligence. Field engineers have to go on-site, drive or walk miles along the power line, find clues to locate the fault spot, and collect evidences for the root cause. Being investigated in power engineering for decades, automated fault diagnosis is identified as one of the key technical challenges in Smart Grid as well [2].

Efforts in the fault diagnosis area include establishing relations between environmental factors and fault causes. For example, Guikema *et al*. [3] and Bernardi *et al*. [4] built statistical models to relate fault to tree trimming and lightning; Xu *et al*. [5] and Sahai *et al*. [6] characterized and modeled features related to tree-caused and animal-caused faults by logistic regression and simple Bayesian networks. Several automated fault diagnosis methods have been proposed by Xu, Chow *et al*. [7-9]. They formulated distribution fault cause identification as a classification problem and applied various biologically inspired

algorithms, including neural networks, fuzzy systems and Artificial Immune Recognition Systems (AIRS).

Instead of relying on expert experience, automated distribution fault diagnosis is generally based on the analysis of historical fault events. Most utilities keep recording faults in their systems and have accumulated large fault databases. Currently, in a typical outage management database, one fault is associated with tens of fields concerning its properties, such as time of occurrence, duration, affected customers, etc. Some of the fields are recorded for different tracking and analyzing purposes that might not be relevant to the root cause. Features used in fault diagnosis, such as seasons, weather conditions, and protective device types, are often chosen by distribution engineers [7].

Unfortunately, human-based feature selection is not easy to generalize. Utilities can record faults in very different manners. Even in the same utility database, fault record fields change over time. For instance, in Progress Energy Carolinas outage database, a field indicating major weather events was added in 2003 to accommodate the new IEEE standard of reliability indices; several more fields specifying actual fault locations were added in 2005 due to the expansion of information collection capability. On the other hand, significant factors change from region to region. Strong wind is an important factor leading to tree-caused faults in wooded areas but may not be as significant in metropolitan areas. Experts cannot be familiar with all the systems in different locations. They need time to gain experience in a new region to select proper features for diagnosing faults locally. Therefore, methods helping select features based on data are necessary, especially when we are able to collect more and more data with the emergence of smart grid technologies.

This paper reviews two commonly used feature selection methods: a) hypothesis test, b) stepwise regression, and introduces another two feature selection methods: c) stepwise selection by Akaike's Information Criterion (AIC), and d) LASSO/ALASSO. These methods select significant features from candidate factors based on data. Advantages and limitations of each method are discussed based on our test on real-world datasets. The rest of this paper is structured as follows: Section II gives the problem formulation; Section III introduces the fundamentals of stepwise selection by AIC and feature selection by LASSO/ALASSO, and compares these four methods; Section IV is case studies of feature selection using these methods on Progress Energy Carolinas outage database; Section V gives the conclusion.

II. PROBLEM FORMULATION

A. Fault Diagnosis as a Classification Problem

Distribution fault diagnosis is formulated as a classification problem in our study. Each fault event can be represented as a vector $[c, f_1, \cdots, f_M]^T$. The classification variable c takes nominal value indicating the root causes of interest, such as tree and animal. The quantitative variables $\mathbf{f} = [f_1, \cdots, f_M]^T$ are features associated with this fault event. Fault diagnosis process can be described as: given historical fault events with known root causes (training data) $\{[c_1, \mathbf{f}_1], [c_2, \mathbf{f}_2], \cdots, [c_N, \mathbf{f}_N]\}$, find a mapping (classifier) $C : \mathbf{f} \to c$ that can map any value of $\mathbf{f}$ to the true value of c. In this paper, we will focus on the discussion of the two-class (binary) problem where c is a binary variable.

B. Feature Selection

As discussed in Section I, some components of the feature vector $\mathbf{f}$ may contain information that is redundant or irrelevant to the root cause. Thus, instead of using all the M factors for fault diagnosis, a new feature vector $\tilde{\mathbf{f}}$ with m ($m \leq M$) components is included in the classification. Those m components are a subset of the M available factors.

Features are selected based on the information they could provide for the classification. Theoretically, there is an optimal subset of the candidate factors for a certain classification problem. However, it is impractical to find this optimal subset when M is large. Using different criteria to determine how significant a candidate factor is, various feature selection methods have been developed.

C. Logistic Regression as a Classifier

Logistic Regression (LR) is a well known statistic method to analyze problems with binary dependent variables [10]. LR builds a model to predict the logarithm odds for the sample to be in one class.

Suppose the two classes under study are $c \in \{0,1\}$, the logistic regression model is:

$$\text{logit}(c=1) = \ln \frac{P(c=1)}{P(c=0)} = \alpha + \boldsymbol{\beta}^T \mathbf{f}\,, \tag{1}$$

where α and $\boldsymbol{\beta}$ are unknown parameters to be estimated by minimizing the negative log-likelihood function:

$$-\log[L(\boldsymbol{\beta})] = -\frac{1}{N}\sum_{i=1}^{N} c_i(\alpha + \boldsymbol{\beta}^T \mathbf{f}_i) - \log(1 + e^{\alpha + \boldsymbol{\beta}^T \mathbf{f}_i})\,, \tag{2}$$

where N is the sample size of the training set.

Equation (1) can be easily solved for the probability of the observed sample being in the class of interest:

$$P(c=1)=\frac{1}{1+e^{-\alpha-\boldsymbol{\beta}^T\mathbf{f}}}. \tag{3}$$

The class is then assigned by comparing the calculated probability with a predefined threshold, naturally 0.5 in a two-class problem. An optimal threshold to minimize the classification error can be found through experiments [7].

LR is used as the classifier in this paper to illustrate the feature selection algorithms. As a model based method, LR is efficient and the reported probability of being one class is intuitive. Nonparametric methods, such as SVM and clustering algorithms, are considered to be more flexible and powerful. However, selecting features based on these methods is more involved in statistics and less easy to explain from engineering perspectives. Thus, we will limit our discussion on feature selection algorithms to LR.

III. Statistical Feature Selection Methods

A. Hypothesis Test and Stepwise Regression

As a common practice, *t*-test is performed for each estimated coefficient when fitting data to a logistic model [10]. A *P*-value is reported by most software to indicate the statistical significance of the estimated coefficient. Thus, a straightforward way of feature selection is to select those significant factors with a small enough *P*-value (usually less than 0.05). A new model only using the significant factors is then estimated again with the same training data.

Stepwise regression has been used for feature selection for long time and is the standard procedure in some statistical software. Features are selected by their discriminant power with controlling the effects of features already in the model. This can be measured by the partial *R*-square value, which is a measurement of the marginal contribution of one factor when all others are already included in the model [11].

Features can be selected in a forward or backward manner. Forward selection starts with no factor and the one providing the best discriminant power enters the model at each step. Backward selection starts with all the factors in the model and removes the one with the least discriminant power each time. Either approach results in a list of factors ordered by their discriminant power. The associated partial *R*-square value gives a quantitative measure of its discriminant power. A preset threshold prevents variables that are not significant from entering the final model [11].

B. Stepwise Selection by Akaike's Information Criterion

Akaike's Information Criterion is a measure of the goodness of fit of an estimated statistical model, defined as

$$AIC = -2\log[L(\boldsymbol{\beta})] + 2k\,, \tag{4}$$

where $-\log[L(\boldsymbol{\beta})]$ is the negative log-likelihood function and k is the number of variables in the model [12].

Intuitively, AIC is a measure of the trade-off between model complexity and accuracy. Generally, a larger likelihood value can be achieved by a model with more variables and the term $2k$ is added to penalize the newly added variables. A minimum AIC can be found when an optimal trade-off is achieved.

Unlike hypothesis test, there is no absolute reference value for AIC to compare to. It is a relative measure to compare several different models on the same dataset. When applied to feature selection, models with different feature combinations are fitted in a stepwise manner and AIC is evaluated. The best feature set is the one yields the minimum AIC.

C. LASSO/ALASSO

The LASSO is a shrinkage and selection method for linear regression proposed by Tibshirani [13]. It minimizes the usual negative log-likelihood with a shrinkage penalty:

$$-\log[L(\boldsymbol{\beta})]+\lambda\sum_{j=1}^{k}J(\left|\beta_j\right|). \tag{5}$$

A common penalty function, so called L_1 penalty, is $J(\left|\beta_j\right|)=\left|\beta_j\right|$. This penalty can lead to an estimated model with smaller mean squared errors than the ordinary least square estimates. In addition, it shrinks small coefficients β_j to exactly zeros and therefore selects those variables with non-zero coefficients.

ALASSO stands for adaptive LASSO, an improved LASSO method by Zuo, Zhang and Lu [14, 15]. The penalty term in ALASSO is a weighted L_1 penalty defined as:

$$\lambda\sum_{j=1}^{k}\frac{\left|\beta_j\right|}{\left|\tilde{\beta}_j\right|}, \tag{6}$$

where $\tilde{\boldsymbol{\beta}}$ is the coefficient vector estimated by the ordinary maximum likelihood method, whose absolute value reflects the relative importance of the factor. There are several nice theoretical properties of ALASSO [14] and it is shown to yield better feature selection results [15].

Parameter λ needs to be tuned in both LASSO and ALASSO. A systematic approach for calculating a series of λ is proposed and implemented by Friedman and Hastie [16]. The best model can be chosen according to Bayesian Information Criterion (BIC).

D. Comparison of Methods

Selecting features by hypothesis test is an effective method with a small number of candidate factors. When the candidate pool is large, the fitted model is heavily affected by the 'noisy' factors and the significant factor set tends to be smaller than enough.

Stepwise regression does give results that make some sense, but it is criticized for reporting the incorrect statistical values, such as partial *R*-square and *P*-value [17]. The serious flaws in its theoretical foundations prevent us using it in the following case studies.

As a summary, these feature selection methods can be compared from the following perspectives:

- Model: all methods apply to generalized linear models.
- Assumption: all of these four methods do not require extra conditions beyond the generalized linear model.
- Computational cost: hypothesis test only fits the model once so it is the fastest among these methods. LASSO/ALASSO needs to minimize the penalized negative log-likelihood function with different λ values thus requires a little more computation. Stepwise selection by AIC is computationally expensive because a large number of feature combinations need to be explored.

- Limitations: hypothesis test is only effective when the candidate pool is small. Stepwise selection by AIC works best when the number of candidate factors is much smaller than the sample size.

IV. CASE STUDIES

Outage database from Progress Energy Carolinas is used in our case study. This dataset contains records of sustained outages in their distribution systems between 2002 and 2006. Of all the outages caused by distribution system faults (in contrast to planned outages and customer side outages), two major causes are of our interest – tree and animal. Other fault causes, such as equipment failures and human interferes, are grouped as one single class.

A. Case 1: Select Features in Three Different Regions

This case study is to demonstrate how feature selection algorithms work with different datasets. Three typical regions in North Carolina are chosen – service region of Ashville operation center for mountainous area, Garner operation center for piedmont area and Wilmington operation center for coastal plains (shown in Fig. 1). Fault characteristics are quite different in these regions. As a direct indication shown in Fig. 2, the percentage of tree-caused faults varies a lot – as high as 46% in Asheville and as low as 13% in Wilmington.

Raw data from the outage database are first cleaned. Outage records with missing values or values that are not defined are deleted. Noises are difficult to be filtered because there is little information about how the data were collected.

Fig. 1. Regions under study.

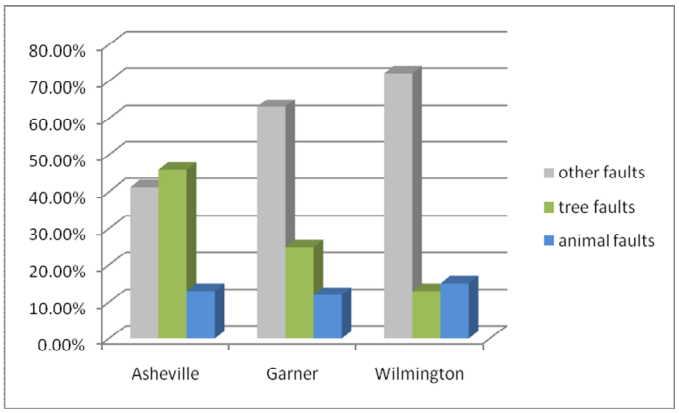


Fig. 2. Percentage of fault causes in three study regions.

Six candidate factors are extracted from the raw data:

PHASE (number of phases affected) = {1,2,3};

SEASON (season) = {spring, summer, fall, winter}, each season includes three months and spring starts at March 1st;

TIME (time of the day) = {morning, afternoon, evening, night}, each period includes six hours and morning starts at 6:00am;

DEVICE (protective device activated) = {break/switch, customer, fuse, recloser, sectionalizer, service transformer, source};

WEATHER (weather conditions when the fault occurred) = {clear weather, extreme temperature, raining, snow/ice, thunderstorm, tornado, windy};

OH_UG (overhead or underground device) = {OH, UG}.

These factors are potentially significant to identify the root cause and are examined using the methods previously introduced. Features selected are shown in Table I and II.

Categorical factors are first converted into dummy variables. For example, PHASE2 and PHASE3 represent the original factor PHASE with value 2 and 3. If PHASE3 is selected for tree-caused faults as in Asheville, it means three-phase is an important indication of tree-caused faults when compared to single-phase. The numbers listed in the tables are the number of variables being selected. The total number of variables is 19 in this case. None of these methods select all the candidate variables. Intuitively, there are some insignificant variables. For instance, fall is not very different from spring in terms of effects on tree-caused faults because the vegetation conditions and weather patterns are very similar. So SEASON3 (fall) should not be selected when SEASON1 (spring) is used as the basis. The algorithms do reject SEASON3 in most cases.

As we expected, different features are significant for identifying different root causes. For example, DEVICE3 (fuse) is generally significant for animal-caused faults but not for tree-caused faults. In different regions, features selected to identify the same root cause are quite different, too. In the same region, hypothesis test always yields the smallest model; ALASSO produces smaller models than LASSO but the features ALASSO selected are not subsets of those selected by LASSO.

TABLE I

SELECTED FEATURES FOR TREE-CAUSED FAULTS IDENTIFICATION IN THREE REGIONS

	Hypothesis Test	***AIC***	***LASSO***	***ALASSO***
Asheville	13	17	18	15
	PHASE3 SEASON2 TIME2/3/4 DEVICE2/7 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 SEASON2/3/4 TIME2/3/4 DEVICE2/3/7 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 SEASON2/3/4 TIME2/3/4 DEVICE2/5/6/7 WEATHER2/4/5/6/8 OH_UG1	PHASE3 SEASON2 TIME2/3/4 DEVICE2/5/6/7 WEATHER2/4/5/6/8 OH_UG1
Garner	9	14	15	9
	SEASON2 DEVICE2/3/7 WEATHER4/5/6/8 OH_UG1	PHASE2 SEASON2 TIME2/4 DEVICE2/3/5/7 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 SEASON2/4 TIME2/4 DEVICE2/6/7 WEATHER2/4/5/6/8 OH_UG1	SEASON2 DEVICE2/7 WEATHER2/4/5/6/8 OH_UG1
Wilmington	9	13	11	14
	SEASON2/4 TIME2/3 DEVICE7 WEATHER4/6/8 OH_UG1	PHASE3 SEASON2/4 TIME2/3 DEVICE2/5/6/7 WEATHER4/6/8 OH_UG1	SEASON2/4 TIME2/3/4 DEVICE3/7 WEATHER4/6/8 OH_UG1	PHASE3 SEASON2/4 TIME2/3 DEVICE2/5/6/7 WEATHER4/5/6/8 OH_UG1

TABLE II

SELECTED FEATURES FOR ANIMAL-CAUSED FAULTS IDENTIFICATION IN THREE REGIONS

	Hypothesis Test	***AIC***	***LASSO***	***ALASSO***
Asheville	14	17	18	17
	PHASE2 SEASON2 TIME2/3/4 DEVICE2/3/7 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 SEASON2/4 TIME2/3/4 DEVICE2/3/5/7 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 SEASON2/3/4 TIME2/3/4 DEVICE2/3/5/7 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 SEASON2/4 TIME2/3/4 DEVICE2/3/5/7 WEATHER2/4/5/6/8 OH_UG1
Garner	8	13	15	12
	PHASE3 TIME2/3 DEVICE3 WEATHER4/6/8 OH_UG1	PHASE2/3 TIME2/3/4 DEVICE2/3 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 SEASON2/3 TIME2/3 DEVICE3/5/7 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 TIME2/3 DEVICE2/3 WEATHER2/4/5/6/8 OH_UG1
Wilmington	7	16	16	10
	SEASON2 TIME2/3 DEVICE3 WEATHER4/6/8 OH_UG1	PHASE2/3 SEASON2/3 TIME2/3/4 DEVICE2/3/5 WEATHER2/4/5/6/8 OH_UG1	PHASE2/3 SEASON2/3 TIME2/3 DEVICE2/3/5/6 WEATHER2/4/5/6/8 OH_UG1	TIME2/3 DEVICE3/5 WEATHER2/4/5/6/8 OH_UG1

B. Case 2: Select Features from Extended Outage Records

This case study presents feature selection with a more complex candidate factor pool. An extended outage dataset for an area around the city of Raleigh is studied (see Fig. 3).

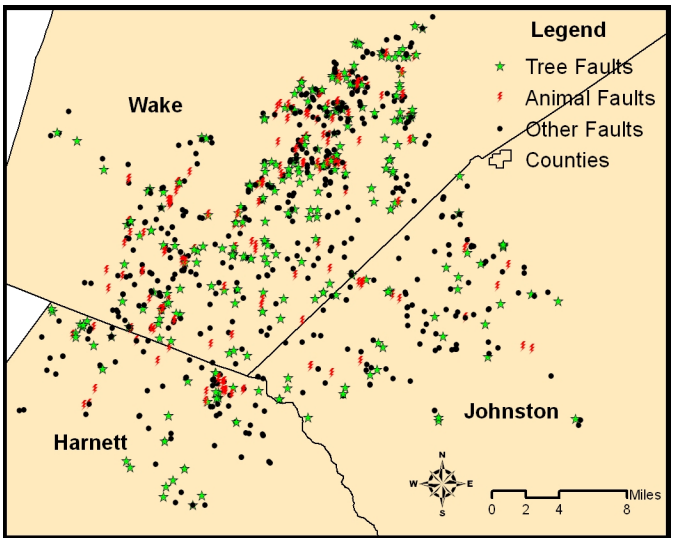


Fig. 3. Region and faults in case 2.

Besides the aforementioned six categorical factors, another eight factors are obtained from other data sources. Obviously erroneous data, such as negative wind speeds and a jumping value in consecutive hourly temperature readings are removed. These new factors are linked to the outage records by spatial or spatial-temporal relations [18], listed as follows:

LULC (land use and land cover type) = {water, developed, barren, forested upland, shrubland, non-natural woody, herbaceous upland, planted/cultivated, wetlands} [19];

DIST_TREE (distance to the nearest trees);

DIST_ROAD (distance to the nearest major roads);

MAX_TEMP (daily maximum temperature);

MIN_TEMP (daily minimum temperature);

PRECIP (daily precipitation);

AVG_WIND (hourly average wind speed);

MAX_WIND (hourly maximum wind speed).

Features selected for tree-caused/animal-caused faults are listed in Table III. With the total number of variables 33, selected features are only around one third of the candidates. Hypothesis test still produces the smallest feature set and features selected for tree-caused and animal-caused faults are quite different.

TABLE III

SELECTED FEATURES FOR TREE-CAUSED AND ANIMAL-CAUSED FAULTS WITH EXTENDED OUTAGE RECORDS

	Hypothesis Test	***AIC***	***LASSO***	***ALASSO***
Tree	8	17	11	10
	SEASON2 WEATHER4/8 OH_UG1 LULC8 DIST_ROAD DIST_TREE AVG_WIND	PHASE3 SEASON2/3 TIME2/3 DEVICE7 WEATHER4/6/8 OH_UG1 LULC3/7/8/9 DIST_ROAD DIST_TREE AVG_WIND	SEASON4 TIME4 DEVICE7 WEATHER4/8 OH_UG1 DIST_ROAD DIST_TREE MIN_TEMP AVG_WIND MAX_WIND	SEASON2/4 DEVICE7 WEATHER2/4/8 OH_UG1 LULC3/5/9
Animal	5	11	14	13
	TIME2 WEATHER4/8 OH_UG1 LULC7	TIME2/3 DEVICE3/7 WEATHER4/6/8 OH_UG1 LULC7 MAX_TEMP AVG_WIND	SEASON4 TIME2/4 DEVICE3/7 WEATHER4/6/8 OH_UG1 LULC7/9 MAX_TEMP PRECIP AVG_WIND	SEASON4 TIME2 DEVICE3/7 WEATHER2/4/6/8 OH_UG1 LULC3/5/7/9

To test the effectiveness of feature selection, fault diagnosis is performed with selected features on this dataset and the results are compared. Suppose the class we are interested in (e.g. tree fault) is positive class, a confusion matrix can be built from the classification result

by counting number of samples falling into the four cases (Table IV). The most commonly used classification performance measure is accuracy (ACC), defined as

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}. \tag{7}$$

TABLE IV

CONFUSION MATRIX

	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	True Positive (TP)	False Negative (FN)
Actual Negative Class	False Positive (FP)	True Negative (TN)

As shown in Fig. 2, outage data is imbalanced in most cases. With animal faults accounting for only around 10%, high classification accuracy close to 90% could be achieved even when all samples are classified as non-animal, which is misleading. Therefore, two more performance measures are used. Probability of detection (POD):

$$POD = \frac{TP}{TP + FN}, \tag{8}$$

is a measure of how well the positive class is correctly detected, with the ideal value 1 indicating all the samples in positive class are identified as positive. False alarm ratio (FAR):

$$FAR = \frac{FP}{TP + FP}, \tag{9}$$

represents how much we can believe it when a sample is classified as positive class. In a perfect case, FAR equals to 0 which means all samples classified as positive are actually positive.

The dataset is randomly split into training set and testing set with equal size. Logistic models are fitted based on training set and used to classify both the training set and testing

set. Experiments are repeated 30 times and the average performance measures are reported in Table V with standard deviation in brackets.

TABLE V

FAULT DIAGNOSIS PERFORMANCE WITH SELECTED FEATURES ON EXTENDED OUTAGE RECORDS

		Full Model		***Hypothesis Test***		***AIC***	
		Train	Test	Train	Test	Train	Test
Tree	ACC	0.799(0.014)	0.751(0.022)	0.773(0.18)	0.760(0.017)	0.789(0.013)	0.758(0.017)
	POD	0.522(0.048)	0.456(0.055)	0.427(0.051)	0.407(0.058)	0.513(0.039)	0.458(0.052)
	FAR	0.270(0.029)	0.346(0.069)	0.286(0.033)	0.310(0.046)	0.277(0.025)	0.350(0.051)
Animal	ACC	0.862(0.014)	0.824(0.016)	0.828(0.013)	0.827(0.014)	0.849(0.011)	0.839(0.015)
	POD	0.498(0.063)	0.378(0.083)	0.049(0.024)	0.045(0.023)	0.392(0.105)	0.367(0.088)
	FAR	0.354(0.027)	0.509(0.070)	0.366(0.194)	0.383(0.190)	0.408(0.062)	0.427(0.067)

		LASSO		***ALASSO***	
		Train	Test	Train	Test
Tree	ACC	0.779(0.014)	0.763(0.015)	0.759(0.017)	0.741(0.019)
	POD	0.464(0.058)	0.440(0.048)	0.323(0.039)	0.304(0.039)
	FAR	0.284(0.030)	0.321(0.060)	0.279(0.047)	0.293(0.067)
Animal	ACC	0.849(0.014)	0.841(0.015)	0.840(0.013)	0.838(0.017)
	POD	0.388(0.082)	0.374(0.065)	0.310(0.147)	0.295(0.132)
	FAR	0.389(0.045)	0.427(0.082)	0.372(0.077)	0.419(0.106)

Generally, the fault diagnosis performance is far from perfect with small POD (around 0.5) and large FAR (around 0.3) due to data imbalance, although the overall ACC achieves around 0.8. The best performance we obtained is with all the candidate factors on training set. However, the drastic performance drop on the corresponding testing set suggests that this model contains too many features to affect its predictability. The worst feature set is the one selected by hypothesis test for animal-caused faults, which shows basically no detecting capability (POD near 0). Features selected by ALASSO lead to a little worse cause identification than stepwise selection by AIC and LASSO for both tree-caused and animal-caused faults.

C. *Discussion*

Our case studies test feature selection methods with both categorical factors and a mixture of categorical and continuous factors. The results confirm that different features need to be used for different fault causes or in different regions. The results also suggest that our model could be over-fitted and lose predictability without feature selection.

Based on our case studies, we can conclude:

- Hypothesis test tends to select fewer features but could fail to find a valid feature set to identify the cause when the number of candidate factors is large.
- Stepwise selection by AIC and LASSO are comparable in terms of the number of selected features and final diagnosis performance, although selected features are not the same.
- ALASSO selects fewer features than LASSO but features being selected are not a subset of those selected by LASSO.
- Although ALASSO is superior to LASSO in feature selection theoretically, fault diagnosis performances with features selected by ALASSO do not show any advantages.

V. CONCLUSION

Selecting features for identifying the root cause is a critical step in distribution fault diagnosis, especially when we have more and more data available with the advancement of information and sensor technologies, and with the emergent Smart Grids. To cope with the

fast growing databases, feature selection methods from statistical community can provide critical tools for power fault managements.

Four feature selection methods: a) hypothesis test, b) stepwise regression, c) stepwise selection by AIC, and d) LASSO/ALASSO are introduced and compared with real-world datasets in this paper. These methods are all applicable to generalized linear models without extra assumptions. Selecting features by hypothesis test is straightforward and easy to use, but it could fail to find a valid feature set when the number of candidate factors is large. Stepwise regression has fundamental flaws theoretically but still gives some meaningful results. It is an option considering its availability in many software packages. Stepwise selection by AIC is easily understandable but computationally expensive on a large dataset. LASSO/ALASSO has nice theoretical properties and is more involved in statistics.

Practically, stepwise selection by AIC and LASSO select different features but the number of selected features and the final diagnosis performance are similar. ALASSO selects fewer features than LASSO but features being selected are not a subset of those selected by LASSO. In terms of fault diagnosis performances, ALASSO is worse than the former two methods.

In general, there is no one single best method for all situations. However, the following could be used as practical guidelines for choosing feature selection algorithms:

- Hypothesis test – a quick screening tool. Do not use it when there are many candidate factors or some of candidate factors are known to be redundant.
- Stepwise regression – use it if any of the existing software package comes handy. The final selection usually makes sense but do not expect the reported order of

significance to be accurate. The threshold for allowing a candidate factor to enter the model is subjective. Always need to be verified with other methods.

- Stepwise selection by AIC – a comprehensive screening tool for large sample sizes and large candidate pools.
- LASSO/ALASSO – another comprehensive screening tool. Use both stepwise selection by AIC and LASSO/ALASSO on the same dataset to get different suggestions.

Features selected by algorithms always need to be validated by field experts. Those algorithms are meant to help engineers to find out information that might be buried under the massive data rather than produce some features that human cannot understand or explain. For a certain region, the integration of feature sets generated by various methods and experts' insights would yield the best result.

VI. ACKNOWLEDGMENT

The authors gratefully acknowledge the contribution of John W. Gajda and Glenn C. Lampley from Progress Energy Carolinas Inc. for their support on data and field experience.

REFERENCES

[1] R. E. Brown, *Electric Power Distribution Reliability*. New York: Marcel Dekker, 2002.

[2] Electricity Advisory Committee, US Department of Energy "Smart grid: enabler of the new energy economy," [Online]. Available: http://www.oe.energy.gov/DocumentsandMedia/final-smart-grid-report.pdf.

[3] S. D. Guikema, R. A. Davidson, and H. Liu, "Statistical models of the effects of tree trimming on power system outages," *IEEE Trans. Power Del.*, vol. 21, no. 3, pp. 1549-1557, 2006.

[4] M. Bernardi, A. Borghetti, C. A. Nucci, and M. Paolone, "A statistical approach for estimating the correlation between lightning and faults in power distribution systems," in *Proc. 9th Int. Conf. Probabilistic Methods Applied to Power Systems (PMAPS)*, Stockholm, Sweden, 2006.

[5] L. Xu, M.-Y. Chow, and L. S. Taylor, "Analysis of tree-caused faults in power distribution systems," in *Proc. 35th North American Power Symp.*, Rolla, MO, 2003.

[6] S. Sahai and A. Pahwa, "A probabilistic approach for animal-caused outages in overhead distribution systems," in *Proc. 9th Int. Conf. Probabilistic Methods Applied to Power Systems (PMAPS),* Stockholm, Sweden, 2006.

[7] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 53-60, 2006.

[8] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-algorithm," *IEEE Trans. Power Syst.*, vol. 22, no.1, pp. 164-171, 2007.

[9] L. Xu, M.-Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using Artificial Immune Recognition System (AIRS) algorithm," *IEEE Trans. Power Systems*, vol. 22, no.1, pp. 198-204, 2007.

[10] D. W. Hosmer, and S. Lemeshow, *Applied Logistic Regression* 2nd ed. New York: Wiley, 2000.

[11] SAS Online Documentation, [Online]. Available: http://support.sas.com/documentation/onlinedoc/91pdf/index_913.html

[12] H. Akaike, "Maximum likelihood identification of Gaussian autoregressive moving average model," *Biometrika*, vol. 60, pp. 255-265, 1973.

[13] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Statist. Soc.*, *Series B*, no. 58, pp. 267-288, 1996.

[14] H. Zou, "The adaptive LASSO and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, pp. 1418-1429, 2006.

[15] H. H. Zhang, and W. Lu, "Adaptive-LASSO for Cox's proportional hazards model," *Biometrika*, vol. 94, pp. 691-703, 2007.

[16] J. Friedman and T. Hastie, "Regularization paths for generalized linear models via coordinate descent," Department of Statistics, Stanford University 2007.

[17] R. Ulrich, "Summary of discussions on .stats usenet groups," [Online]. Available: http://www.pitt.edu/~wpilib/statfaq/regrfaq.html.

[18] Y. Cai and M.-Y. Chow, "Exploratory analysis of massive data for distribution fault diagnosis in Smart Grids," presented in *IEEE Power & Energy Society General Meeting,* Calgary, Canada, 2009.

[19] USGS Land Cover Institute, "NLCD Land Cover Class Definitions," [Online]. Available: http://landcover.usgs.gov/classes.php.

CHAPTER IV

CAUSE-EFFECT MODELING AND SPATIAL-TEMPORAL SIMULATION OF POWER DISTRIBUTION FAULT EVENTS

Yixin Cai Mo-Yuen Chow

ycai2@ncsu.edu chow@ncsu.edu

Department of Electrical and Computer Engineering

North Carolina State University

Raleigh, NC 27695 USA

This chapter is accepted for publication in IEEE Transactions on Power Systems.

CAUSE-EFFECT MODELING AND SPATIAL-TEMPORAL SIMULATION OF POWER DISTRIBUTION FAULT EVENTS

Abstract-- Modeling and simulation are important tools in the study of power distribution faults due to the limited amount of actual data and the high cost of experimentation. Although a number of software packages are available to simulate the electrical signals, approaches for simulating fault events in different environments have not been well developed. In this paper, we propose a framework for modeling and simulating fault events in power distribution systems based on environmental factors and the cause-effect relationships among them. The spatial and temporal aspects of significant environmental factors leading to various faults are modeled as raster maps and probability functions, respectively. The cause-effect relationships are expressed as fuzzy rules and a hierarchical fuzzy inference system is built to infer the probability of faults in the simulated environments. A test case simulating a part of a typical city's power distribution systems demonstrates the effectiveness of the framework in generating realistic distribution faults. This work is helpful in fault diagnosis for different local systems and provides a configurable data source to other researchers and engineers in similar areas as well.

Index Terms—discrete event simulation, fault diagnosis, fuzzy systems, power distribution faults, power system simulation, system modeling.

I. INTRODUCTION

Distribution faults are the major source of customer reliability problems in power systems and have been drawing more and more attention in recent decades [1]. In addition to studying the electrical characteristics of faults, such as fault location based on monitored voltages and currents [2, 3] and the experimental characterization of electrical signals generated during tree contact [4] and lightning strikes [5], many researchers have treated distribution faults as discrete events and established relationships between environments and the root cause of the faults. To mention some, Xu *et al*. [6] and Sahai and Pahwa [7] characterized and modeled environmental factors related to tree-caused and animal-caused faults by logistic regression and Bayesian networks; Xu and Chow *et al*. [8, 9] used biologically inspired algorithms, including artificial neural networks (ANN), fuzzy systems, and artificial immune recognition systems (AIRS), to identify the fault's root cause based on environmental properties.

Due to the limited amount of actual data and the high cost of experimentation, modeling and simulation are essential tools in studying power distribution faults. Although a number of software packages are available to simulate the electrical signals, the study of faults as discrete events is still limited by the lack of modeling and simulation approaches. The stochastic nature of faults, spatial heterogeneity of distribution systems, and measurement noises make it difficult for one single dataset to represent all systems. Thus, researchers often find that an algorithm performing well on one set of fault records may not achieve the same performance on another fault dataset. With limited access to the same fault dataset, comparing work done by different research groups can be similar to comparing apples and oranges.

In this paper, we propose a framework for modeling and simulating fault events in power distribution systems based on environmental factors and the cause-effect relationships among them. This framework provides a configurable data source capable of representing different local systems in various environments. This data source would help in developing and evaluating distribution fault diagnosis algorithms. Moreover, repeated simulations under different conditions may help engineers to identify regions and time periods with an unusually high probability of faults so that investments can be prioritized and crew deployments can be planned.

Cause-effect relationships have been investigated to help diagnose faults in distribution systems. Deterministic cause-effect relationships are used to locate the fault section. By discovering the causal relationship between the fault section and the corresponding sequence of protective actions, the actual fault section can be identified based on the recorded protective actions. To cope with the complex scenarios of a large distribution system and the inherent noises and uncertainties in the recorded protective actions, Chen *et al*. [10] used a cause-effect network and fuzzy sets, while Zhang *et al*. [11] and Hor *et al*. [12] proposed using Rough Set theory to discover the actual relationships. Reliability-based modeling of equipment failure is also well studied. The residual life of a device before its next failure is modeled as various probability distributions. Techniques to calibrate and validate the model parameters are developed as well [1, 13]. In this type of model, faults are more likely to occur as time advances. Taking the environment into consideration, Kuntz *et al*. [14] and Radmer *et al*. [15] included yearly average temperature and tree growth status in the reliability model via regression and ANN.

These techniques provide good tools for fault analysis and serve their purpose quite well. However, they are not adequate for the cause-effect simulation of fault events. First, distribution system faults are essentially stochastic events without any deterministic causal relationships. The cause-effect model can only represent the approximate trend of faults with large uncertainties. Second, the fault characteristics are location-dependent and vary over time. Faults in a new metropolitan distribution system should behave quite differently from those in a 40-year-old single-phase long feeder in a rural area. Thus, both spatial and temporal information need to be considered in the model. Last but not least, environmental factors, most importantly weather conditions, need to be modeled and simulated with a time resolution of minutes or hours instead of a monthly or yearly average.

All these requirements make the cause-effect modeling and spatial-temporal simulation of distribution fault events a challenging task. In the following sections, we lay out our solution. We first identify significant environmental factors leading to distribution faults, find appropriate models describing these natural phenomena, and simulate the environments at different times and locations. Then the causal relationships between these factors and various fault types are investigated. Knowledge gained from field experience and the analysis of historical data is represented as rules and modeled using fuzzy inference systems (FIS) [16]. As the causal relationships apply to all systems and environmental factors vary across space and time, the generated fault events can represent realistic characteristics under a wide range of conditions.

II. Environmental Modeling and Simulation

The fundamental information needed for the cause-effect modeling and spatial-temporal simulation of distribution faults is about the power system itself, such as its power line topologies and switchgear configurations. Some of these properties affect the fault characteristics (e.g., whether the power line is overhead or underground). Others are kept for post-fault analysis.

The first step in environmental modeling is to model the significant environmental factors that lead to various faults. The selection of significant factors and the modeling of the spatial and temporal aspects of these factors are discussed in this section.

A. Significant Environmental Factors

Field experience and the knowledge gained analyzing historical fault data are used to select the significant environmental factors. For example, it is easy to understand that when strong winds blow the tree branches harder, we are more likely to have tree-caused faults. Meanwhile, strong winds reduce animal activities and historical data tells us that fewer animal-caused faults occur on windy days (as shown in Fig. 1). Therefore, wind speed is chosen as a significant factor. As recommended by experienced distribution engineers, factors identified as significant for fault diagnosis [17], such as season and rainfall intensity, are considered important for environmental modeling in this paper. As environmental characteristics change, more factors may be required to model a particular local system.

B. Modeling Spatial Information

The spatial properties of environments describe the locations of those geographic features of interest, such as the layout of highways and vegetation. Maps are the everyday tool for

representing spatial information and are used in our model. Although power lines are usually represented in vector format in power system software, raster maps are preferred in our model due to their flexibility for modeling various features and their ease of manipulation [18]. A raster map consists of grids of a fixed size, where every grid is given a particular number to represent the value, category, or presence of a certain factor. For example, with other layers containing additional spatial information, land use types are represented as the different integer values of the cells, shown as different colors in Fig. 2.

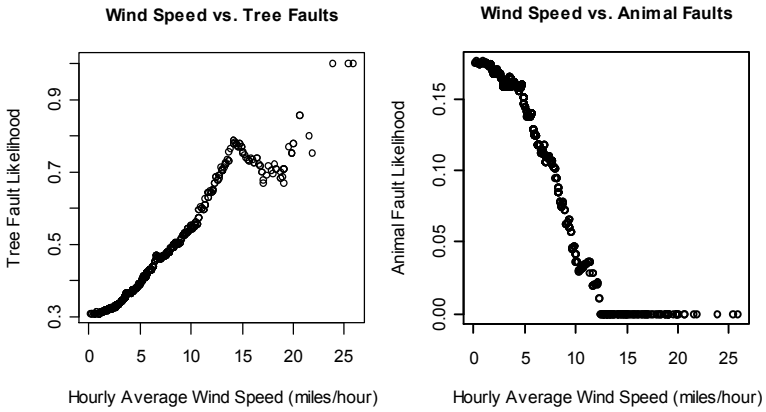


Fig. 1. Likelihood of tree-caused and animal-caused faults under different wind speeds. (Data source: distribution faults in the Garner operation center of Progress Energy Carolinas between 2005 and 2006)

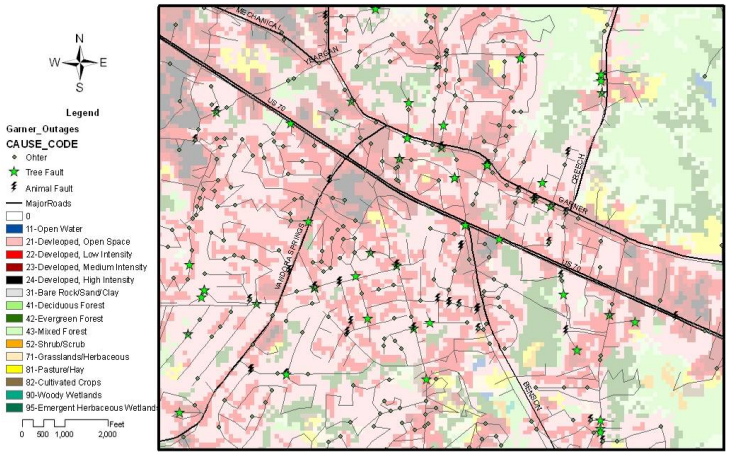


Fig. 2. An example of raster maps showing land use types.

Spatial information is assumed to be static during the simulation period. Although some of the factors do change over time (e.g., the distribution system may expand, urban development may change roads or land use), these changes usually take a long time to complete. Then, they can be treated as new environments, modeled in updated maps, and simulated in a new study period.

C. Modeling Temporal Information

Theoretically, every aspect of the environment is constantly changing at a different rate. For example, utility companies manage the trees along power lines and carry out equipment maintenance and upgrades every few years; the starting and ending dates of a season drift slowly from year to year; precipitation patterns change throughout the year; wind speed changes every instant.

Weather conditions are the most significant temporal information affecting the cause-effect model. Assume the region under simulation is small enough that the weather conditions within it can be regarded as the same. A large region can be divided into smaller areas, modeled, and simulated separately. Three major weather measures, temperature, wind speed, and rainfall intensity, are considered in this paper.

Temperature is widely known to have yearly and daily cyclic patterns, which are modeled as:

$$T = \overline{T} + A\sin\frac{2\pi d}{365} + B\cos\frac{2\pi d}{365} + a\sin\frac{2\pi h}{24} + b\cos\frac{2\pi h}{24} + N(0,1), \qquad (1)$$

where $\overline{T}$ is the yearly average temperature, d is the day of the year (1, 2, …, 365), h is the hour of the day (0, 1, 2, …, 23), A, B, a, and b are parameters reflecting the amplitude and

phase of the sinusoidal waves in (1). One example of the observed hourly average temperature during a three-year period is shown in Fig. 3. Note that even though this model does not provide the exact temperature patterns, it gives an extraction of features that is good enough for the simulation.

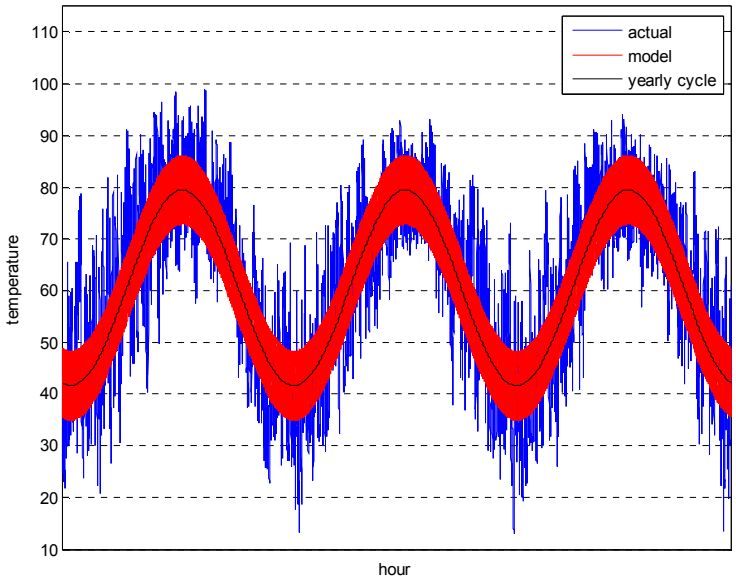


Fig. 3. The 3-year hourly average temperature. (Data source: CLAY weather station in Clayton, NC, between 2004 and 2006)

In the literature, wind speed is usually modeled by a Weibull distribution [19], described as:

$$f(x;\lambda,k)=\begin{cases}\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k} & x\geq 0\\ 0 & x<0\end{cases}, \qquad (2)$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. One example of the actual 1-year hourly average wind speed is shown in Fig. 4.

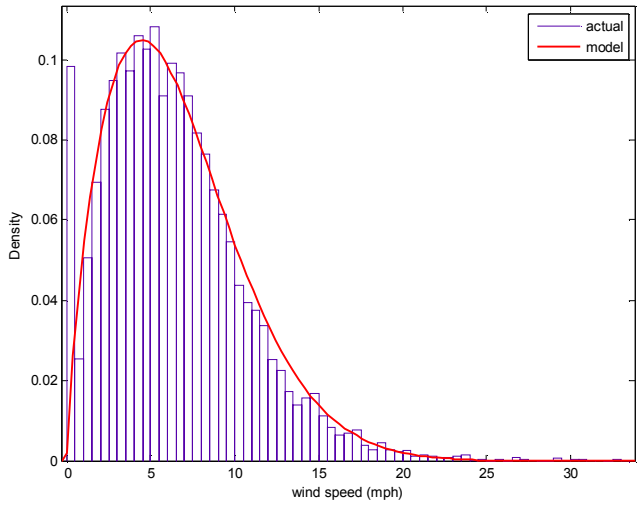


Fig. 4. The 1-year hourly average wind speed. (Data source: LAKE weather station in Raleigh, NC, in 2003)

As shown in Fig. 4, the probability of the wind speed being 0 is much greater than that described in the Weibull model. Thus, we first add the constraint that the probability that the wind speeds are greater than 0 and then use a Weibull distribution to model the speed. Another practical constraint on wind speed is that it cannot be arbitrarily high. Although very large values appear with a tiny probability in the model (2), the upper limit is set to prevent unrealistic wind speeds from being generated. The final wind speed model used in the simulation is as follows:

$$f(x;\lambda,k)=\begin{cases} 0 & x>x_{\max} \\ \dfrac{kp}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x_{\max}\geq x>0 \\ 1-p & x=0 \\ 0 & x<0 \end{cases}, \tag{3}$$

where x is the wind speed, $x_{\max} > 0$ is the maximum wind speed that is allowed, $k > 0$ is the shape parameter, $\lambda > 0$ is the scale parameter, and $p > 0$ is the probability of the wind speed being greater than 0. Based on our observations, p often takes a value close to 1. Note that λ

is not the average wind speed. However, with p and k fixed, λ can be used to represent the average strength of the wind.

Similarly, the intensity of rainfall is modeled as an exponential distribution constrained by the probability of rainfall, as follows:

$$f(x;\lambda)=\begin{cases} 0 & x>x_{\max} \\ \frac{p}{\lambda}e^{-x/\lambda} & x_{\max}\geq x>0 \\ 1-p & x=0 \\ 0 & x<0 \end{cases}, \tag{4}$$

where x is the rainfall intensity, $x_{\max} > 0$ is the maximum rainfall intensity allowed, $\lambda > 0$ is the scale parameter, and $p > 0$ is the probability of rainfall. Unlike wind, rainfall is a relatively rare weather event, so the value of p in (4) is usually set close to 0. One example of the actual 1-year hourly rainfall data is shown in Fig. 5. Note that hours with exactly zero rainfall are excluded from the plot.

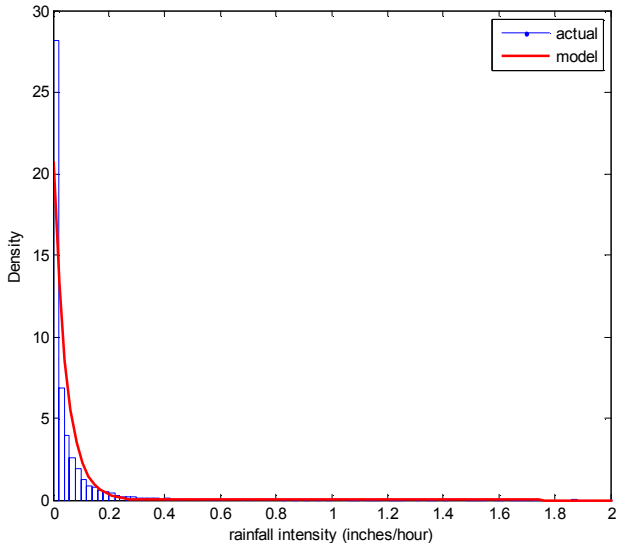


Fig. 5. The 1-year hourly rainfall intensity excluding the hours without any rainfall. (Data source: RDU weather station in Morrisville, NC, in 2005)

III. CAUSE-EFFECT MODELING OF FAULT EVENTS

Cause-effect relationships summarize the typical correlation between environmental factors and faults, disregarding the time and the location. For example, flourishing trees around power lines during a thunderstorm are very likely to lead to tree-caused faults. In our model, these causal relationships are represented by fuzzy rules and are used to infer the probability of various faults given the environmental conditions.

A. Fuzzy Environmental Factors

To apply the fuzzy rule-based inference, significant environmental factors are first converted to fuzzy sets. Membership functions are defined to map the values generated by the environmental models into linguistic descriptions.

For categorical values, the linguistic categories are already defined. Fuzzy sets mainly represent the uncertainty of the input value. The number of fuzzy sets used to represent different categories is the same as the number of levels in the input variable. The membership function for these categorical variables has multiple peaks where the location and peak value can be specified. The uncertainty of the input being one category is described by varying the height of the peaks. For each fuzzy set, the highest peak represents the corresponding input level, while the lower peaks at other values represent the uncertainty. One example uses nine fuzzy sets to represent the nine land use types defined by the USGS Land Cover Institute [20]. Fig. 6 shows the fuzzy membership function used to describe land use type *forest*. For categories that are clearly separable from *forest*, such as *water*, the height of the lower peak is close to 0. For categories that are not so easy to differentiate, such as *non-natural woods*, the lower peak could be of a height similar to the highest one.

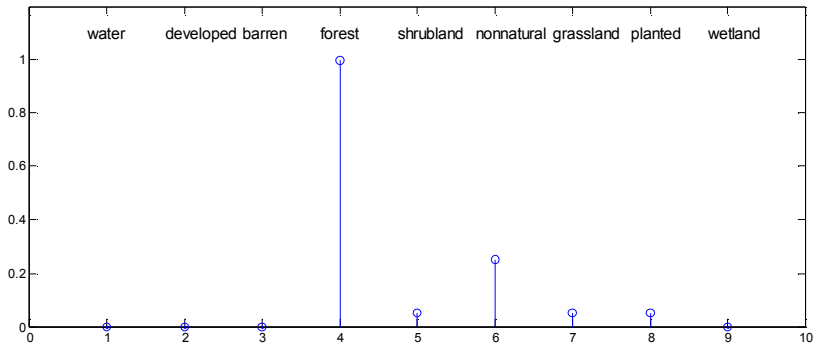


Fig. 6. An example of membership functions for categorical variables.

Continuous values are more conventional as input to fuzzy systems. Therefore, a large collection of standard membership functions, including triangular, trapezoid, and Gaussian, is readily available in the Matlab Fuzzy Logic Toolbox [21]. Based on the nature of the input variables, fuzzy sets with corresponding membership functions are defined. For example, the distance to the nearest trees and to the nearest roads can be represented by three fuzzy sets, *near*, *med*, and *far*. Intuitively, a tree or a road is near a power line if it is within a couple of meters. Considering that the common spatial resolution of raster maps is 30 meters, distances within 30 to 60 meters are defined as *near*, those of more than 100 meters are considered as *far*, and distances in-between are regarded as *med*. As an example we used the standard trapezoid membership function, as shown in Fig. 7.

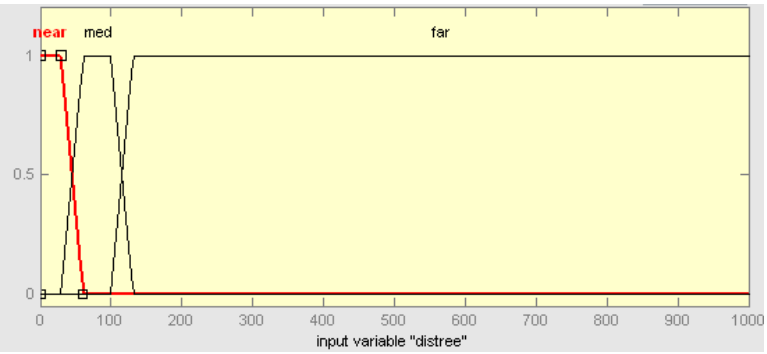


Fig. 7. An example of membership functions for continuous variables.

Fuzzy sets and membership functions are defined to reflect most of the common situations under study. We may need to change a membership function or use a different number of fuzzy sets when modeling different environments. For example, two fuzzy membership functions describing the *dry* and *wet* seasons in southern California may better reflect reality than the four fuzzy sets of *spring*, *summer*, *fall*, and *winter*, which are more suitable in eastern North Carolina.

B. Hierarchical Rule Base

If we use n input variables and each variable has m fuzzy sets, the total number of rules needed in a single FIS is m^n. This number grows exponentially with n. This so-called *curse of dimensionality* makes it very difficult to design and implement a complex FIS. This problem has been addressed extensively in fuzzy system literature and one widely used solution is the hierarchical fuzzy system (HFS) [22, 23]. Breaking the system down into several subsystems can substantially reduce the number of rules needed while the HFS retains the properties of the original FIS [24]. The HFS shown in Fig. 8 (used in the illustrative example in Section IV) has 122 rules, which is less than 0.1% of the number required by a single-layer FIS.

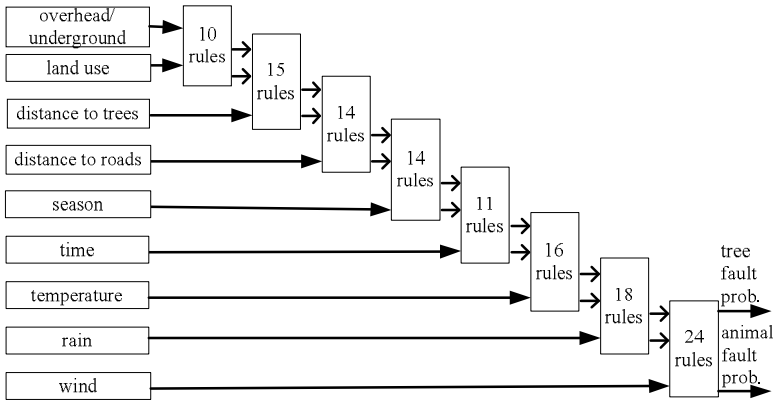


Fig. 8. The structure of a hierarchical fuzzy system (HFS).

As an example, the knowledge regarding underground or overhead lines in the first subsystem is that underground systems are less prone to external interference, thus are less likely to have tree-caused or animal-caused faults. In the form of rules, this is stated as "if the system is underground, then the probability of tree-caused faults is low and the probability of animal-caused faults is low." The following subsystem takes the outputs of the probability of tree-caused or animal-caused faults from the previous subsystem and an additional environmental factor as inputs, and infers new probabilities. Sample response surfaces of the last subsystem regarding wind speed are shown in Fig. 9. As we can see, when the wind speed grows higher, the probability of tree-caused faults increases and the probability of animal-caused faults decreases, reflecting the trends we showed in Fig. 1.

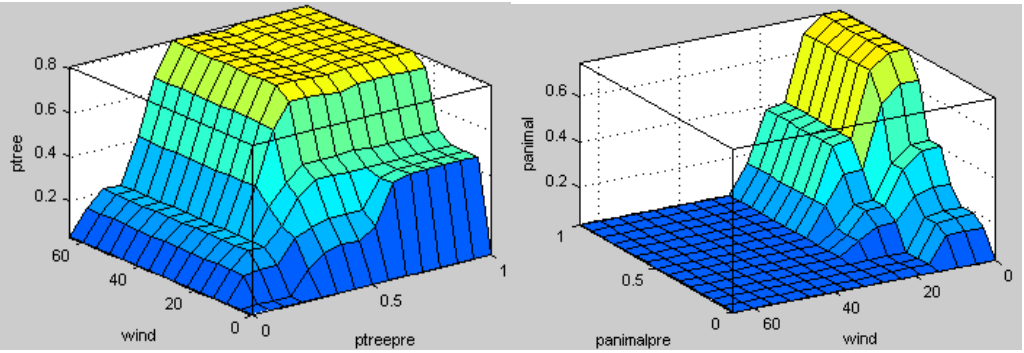


Fig. 9. Sample response surfaces of the input wind speed.

C. Fault Event Generation

All the available environmental factors are evaluated at every map cell where there are valid distribution system components. For each cell, the probability of faults comes from the HFS; fault events are generated according to these probabilities; and generated events with the corresponding properties are returned and stored in a file.

IV. AN ILLUSTRATIVE EXAMPLE

A. Simulator Structure and GUI

To demonstrate the cause-effect modeling and spatial-temporal simulation framework, we have developed a power distribution fault simulator with Matlab [25]. The fault simulator consists of three major parts: input module, fault generator, and graphic user interface (GUI), as shown in Fig. 10.

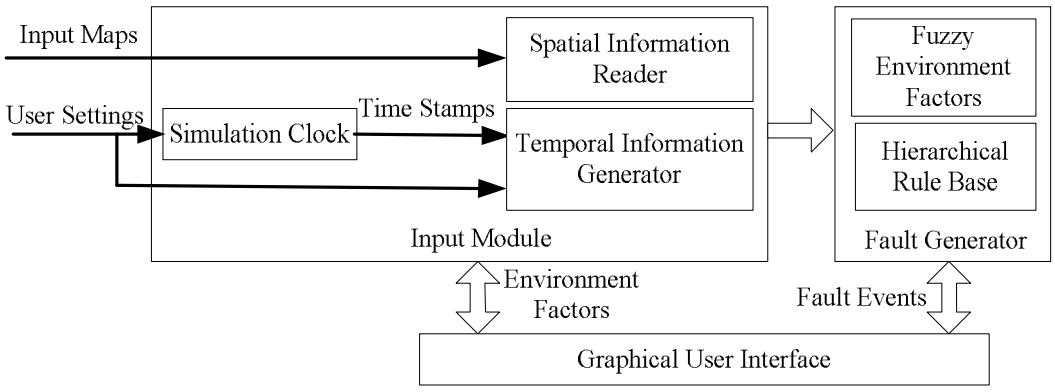


Fig. 10. The overall structure of the fault simulator.

The input module takes user settings in the form of parameters and maps, and converts them to environmental factors. These settings describe the layout and properties of the distribution system under study, the major geographical features, such as land use, vegetation, and roads, and local weather characteristics. The fault generator converts the environmental factors into fuzzy representations, infers the probability of tree-caused and animal-caused faults using the hierarchical fuzzy rules, and generates and records fault events. Users interact with the simulator through the GUI (Fig. 11).

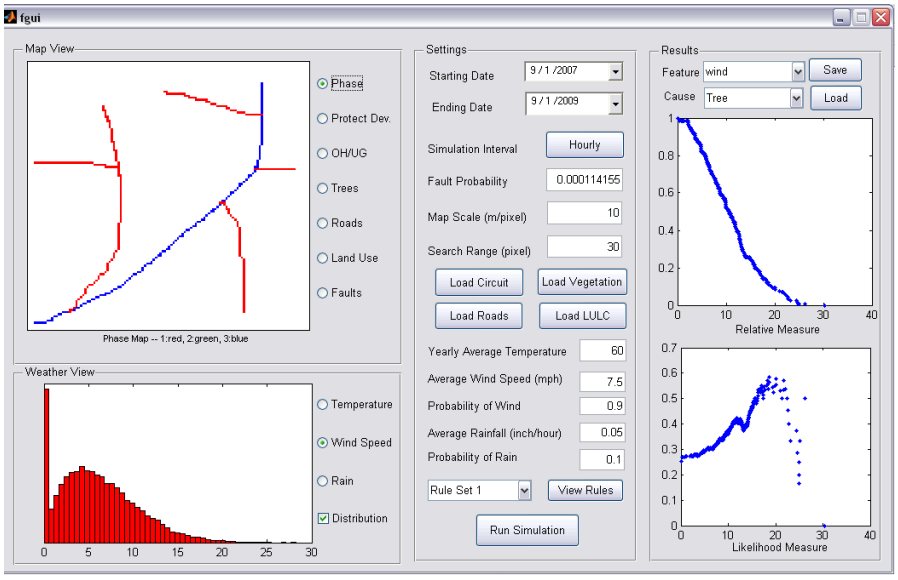


Fig. 11. A sample GUI for the fault simulator.

B. Environmental Factors

Four maps regarding circuits, vegetation, roads, and land use types are required for this simulator. The circuit map describes the distribution system layout with properties encoded in three layers, including the number of phases, the devices protecting the section, and whether it is overhead or underground. The vegetation map records where the tall trees, low bushes, and grasses are. Each category is encoded by different integer values in one layer. The road map represents the layout of major roads in the area under simulation. The land use map encodes the land use type by different integer values in one layer. Fig. 12 shows the spatial information loaded from a set of sample maps. With dense distribution lines, scattered trees, and most land in developed status, this is a typical distribution system for a city region.

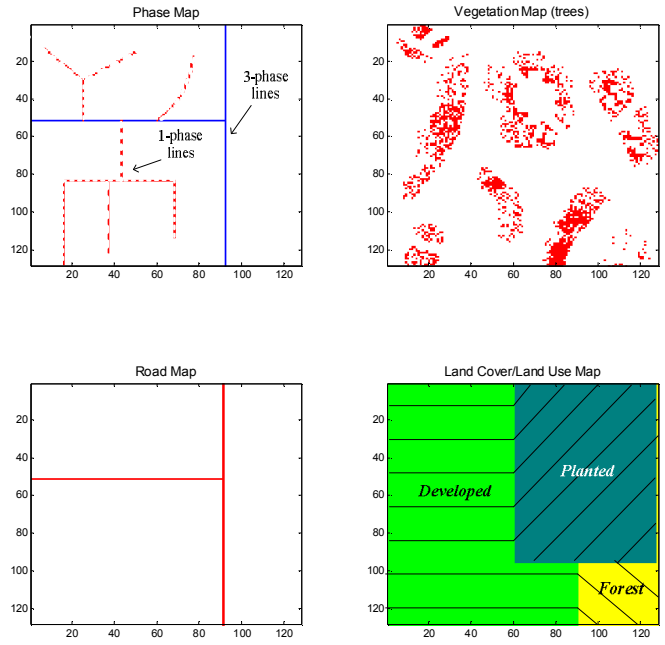


Fig. 12. An example of spatial information loaded from maps.

By fitting the model (1) to the actual data shown in Fig. 3, the following temperature model is used in our simulation:

$$T = 60.5 - 5.0\sin\frac{2\pi d}{365} - 18.3\cos\frac{2\pi d}{365} - 6.4\sin\frac{2\pi h}{24} - 2.2\cos\frac{2\pi h}{24}. \qquad (5)$$

In the wind speed model (3), k is set to be 1.7. λ=7.46 and p=0.95 are estimated from the data shown in Fig. 4. Similarly, λ equals to 0.48 and p equals to 0.11 in the rainfall model (4), as estimated from the data shown in Fig. 5. These models are then used to generate weather conditions during a 2-year period. The histogram of generated wind speed can be found in the bottom-left corner of Fig. 11.

C. Fuzzy Rules

Nine environmental factors are considered to be significant in the simulation, including overhead or underground power lines, land use, distance to trees, distance to major roads,

season, time of day, temperature, wind speed, and rainfall intensity. They are organized as an HFS, as shown in Fig. 8.

The two categorical factors, underground or overhead and land use, are fuzzified based on the membership functions shown in Fig. 6. Fuzzy sets for the distance to trees and major roads are as shown in Fig. 7.

The definition of season and time of day is inherently fuzzy. Although spring begins generally in March in North America, the exact date varies in different locations and from year to year. Similarly, in different places, different seasons, and for different purposes, morning could start as early as 5:00am, or as late as 8:00am. Considering the cyclic property of the hours of the day and the days of the year, a multi-modal trapezoid membership function is used for *evening*, *night*, and *winter*, as shown in Fig. 13.

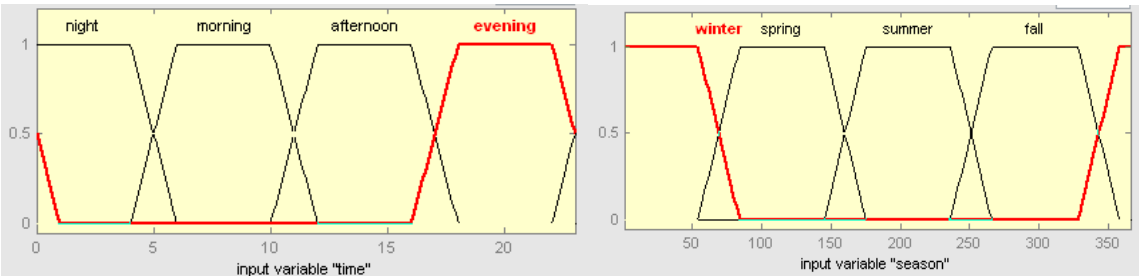


Fig. 13. Membership functions for the season and time of day.

Weather factors are fuzzified based on our daily description of the weather. Temperature is represented by five fuzzy sets, *extrecold*, *cold*, *normal*, *hot* and *extrehot*. The normal temperature ranges from about $55^{\circ}F$ to $85^{\circ}F$. Temperatures above $100^{\circ}F$ are extremely hot and those below $20^{\circ}F$ are extremely cold. The intensity of rainfall is usually described in six levels [26]. We adopt the scale and simplify it to *norain*, *light*, *med*, and *heavy*. Similarly, wind speed is represented by *nowind*, *light*, *med*, and *strong*, according to the commonly

used Beaufort scale [27]. All these fuzzy sets use the standard trapezoid membership functions similar to those in Fig. 7.

As shown in Fig. 14, the same trapezoid membership functions are used for the probability of tree-caused and animal-caused faults, both in the final output and all the intermediate results. The probability of animal-caused faults at *med*, *high*, and *veryhigh* level is slightly lower than those of tree-caused faults, as observed in the actual data.

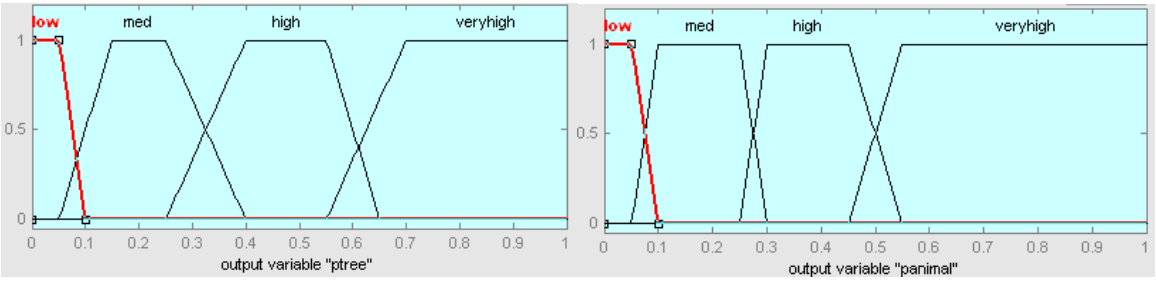


Fig. 14. Membership functions for the probability of tree-caused and animal-caused faults.

D. A Test Case

The simulation settings for the test case are summarized in Table I. It is to simulate a part of the actual distribution systems in the town of Garner, North Carolina.

TABLE I

SIMULATION SETTINGS FOR THE TEST CASE

Simulation period	01/01/2005~12/31/2006
Time step	1 hour
Maps	as shown in Fig. 12
Map resolution	30 meters
Map extent	3840 meters by 3840 meters
Temperature parameters	$\bar{T}$ = 60.5, A = -5.0, B = -18.3, a = -6.4, b = -2.2
Wind parameters	k = *1.7*, λ = 7.46, p = 0.95
Rainfall parameters	λ = 0.48, p = 0.11
Environmental factors	Overhead/underground lines, distance to trees, distance to major roads, season, time, temperature, rainfall intensity, wind speed
No. of subsystems in the FIS	8
No. of rules in the FIS	122

The actual fault data are the power outage records in the distribution system of Progress Energy Carolinas within Garner operation center between 2005 and 2006. We have integrated real-time weather measurements and other geographical information with the outage records using spatial and spatial-temporal relationships [28], so the actual measurements of the environmental factors while the outages occurred are available.

The simulation ran for around 270 minutes on a laptop with a 2.0GHz dual-core processor and 2GB RAM, and generated 131 tree-caused and 137 animal-caused fault events. As a tool to represent fault characteristics, the likelihood of tree-caused and animal-caused faults under selected environmental factors is shown in Fig. 15.

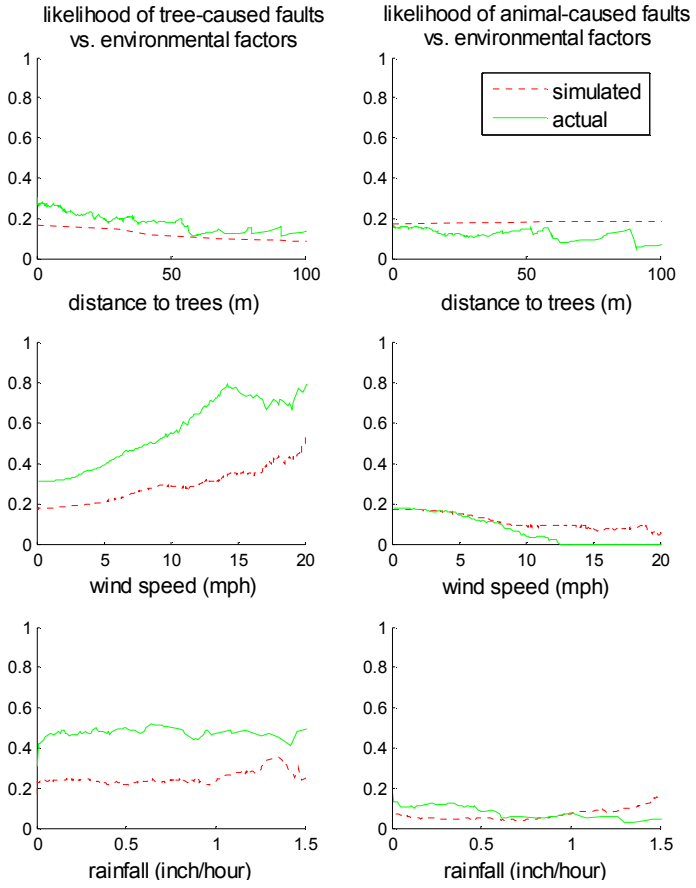


Fig. 15. Likelihood measures of selected environmental factors.

As we can see in Fig. 15, the trend of simulated fault likelihood matches the actual situation in general. Note that we have a relatively small sample of faults with both the simulated and the actual data. This leads to large variations when estimating the fault likelihood, so an exact match of fault likelihood is hardly achievable.

V. CONCLUSION

In this paper, a framework of cause-effect modeling and spatial-temporal simulation of power distribution fault events is proposed to cope with the data availability problem faced by many researchers in the field of distribution fault diagnosis.

Significant environmental factors leading to distribution faults are first identified based on the understanding of distribution faults and the analysis of historical fault data. The spatial information of these factors is modeled as raster maps. The temporal information of weather conditions is modeled as probability functions. The causal relationships among these factors and various types of faults are investigated and expressed in the form of fuzzy rules. A hierarchical fuzzy inference system is built to infer the probability of tree-caused and animal-caused faults in the simulated environments. Fault events are generated and recorded accordingly. The test case demonstrates the effectiveness of the simulation in generating realistic distribution faults.

By changing the environmental settings, we are able to generate fault events for different local systems during different time periods. This framework would help develop and evaluate fault diagnosis algorithms and predict regions with high probability of faults during a certain

time period. It also provides a configurable data source as a useful tool for other researchers and engineers in similar areas.

VI. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of John W. Gajda and Glenn C. Lampley from Progress Energy Carolinas Inc. for their support on data and field experience.

REFERENCES

[1] R. E. Brown, *Electric Power Distribution Reliability*. New York: Marcel Dekker, 2002.

[2] S.-J. Lee, M.-S. Choi, S.-H. Kang, B.-G. Jin, D.-S. Lee, B.-S. Ahn, N.-S. Yoon, H.-Y. Kim, and S.-B. Wee, "An intelligent and efficient fault location and diagnosis scheme for radial distribution systems," *IEEE Trans. Power Del.*, vol. 19, no. 2, pp. 524–532, Apr 2004.

[3] 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.

[4] J. A.Wischkaemper, C. L. Benner, and B. D. Russell, "Electrical characterization of vegetation contacts with distribution conductors—Investigation of progressive fault behavior," in *Proc. IEEE PES Transmission and Distribution Conf. Expo.*, Chicago, IL, 2008.

[5] T. Miyazaki and S. Okabe, "A detailed field study of lightning stroke effects on distribution lines," *IEEE Trans. Power Del.*, vol. 24, no. 1, pp. 352–359, Jan 2009.

[6] L. Xu, M.-Y. Chow, and L. S. Taylor, "Analysis of tree-caused faults in power distribution systems," presented at the 35th North American Power Symp., Rolla, MO, 2003.

[7] S. Sahai and A. Pahwa, "A probabilistic approach for animal-caused outages in overhead distribution systems," presented at the 9th Int.Conf. Probabilistic Methods Applied to Power Systems (PMAPS), Stockholm, Sweden, 2006.

[8] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp.164–171, Feb 2007.

[9] L. Xu, M.-Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using artificial immune recognition system (AIRS) algorithm," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 198–204, Feb 2007.

[10] W.-H. Chen, C.-W. Liu, and M.-S. Tsai, "On-line fault diagnosis of distribution substations using hybrid cause-effect network and fuzzy rule-based method," *IEEE Trans. Power Del.*, vol. 15, no. 2, pp. 710–717, Apr 2000.

[11] Q. Zhang, Z. Han, and F. Wen, "A new approach for fault diagnosis in power systems based on rough set theory," presented at the 4th Int. Conf. Advances in Power System Control, Operation and Management, Hong Kong, China, 1997.

[12] C.-L. Hor, P. A. Crossley, and S. J. Watson, "Building knowledge for substation-based decision support using rough sets," *IEEE Trans. Power Del.*, vol. 22, no. 3, pp. 1372–1379, Jul 2007.

[13] R. E. Brown, G. Frimpong, and H. L. Willis, "Failure rate modeling using equipment inspection data," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 782–787, May 2004.

[14] P. A. Kuntz, R. D. Christie, and S. S. Venkata, "Optimal vegetation maintenance scheduling of overhead electric power distribution systems," *IEEE Trans. Power Del.*, vol. 17, no. 4, pp. 1164–1169, Oct 2002.

[15] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, and R. H. Fletcher, "Predicting vegetation-related failure rates for overhead distribution feeders," *IEEE Trans. Power Del.*, vol. 17, no. 4, pp. 1170–1175, Oct 2002.

[16] T. J. Ross*, Fuzzy Logic with Engineering Applications*, 2nd ed. New York: Wiley, 2004.

[17] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Statistical feature selection from massive data in distribution fault diagnosis," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 642-648, May 2010.

[18] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind*, Geographic Information Systems and Science*. Chichester, U.K.:Wiley, 2001.

[19] Wind Power. [Online]. Available: http://en.wikipedia.org/wiki/Wind_power.

[20] USGS Land Cover Institute, NLCD Land Cover Class Definitions. [Online]. Available: http://landcover.usgs.gov/classes.php.

[21] Mathworks, Fuzzy Logic Toolbox User's Guide. [Online]. Available: http://www.mathworks.com/access/helpdesk/help/toolbox/fuzzy/.

[22] R. R. Yager, "On the construction of hierarchical fuzzy systems models," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 28, no. 1, pp.55–66, Feb 1998.

[23] M.-L. Lee, H.-Y. Chuang, and F.-M. Yu, "Modeling of hierarchical fuzzy systems," *Fuzzy Sets Syst.*, vol. 138, no. 2, pp. 343–361, Sep 2003.

[24] L.-X. Wang, "Analysis and design of hierarchical Fuzzy systems," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 5, pp. 617–624, Oct 1999.

[25] Y. Cai and M.-Y. Chow, "Cause-effect modeling and simulation of power distribution fault events," in *Proc. IEEE Power & Energy Soc. General Meeting*, Minneapolis, MN, 2010.

[26] Rain. [Online]. Available: http://en.wikipedia.org/wiki/Rain.

[27] Wind. [Online]. Available: http://en.wikipedia.org/wiki/Wind.

[28] Y. Cai and M.-Y. Chow, "Exploratory analysis of massive data for distribution fault diagnosis in Smart Grids," in *Proc. IEEE Power & Energy Soc. General Meeting*, Calgary, AB, Canada, 2009.

CHAPTER V

A NOVEL SAMPLING STRATEGY FOR DISTRIBUTION FAULT DIAGNOSIS: SMALL WORLD STRATIFICATION

Yixin Cai ycai2@ncsu.edu

Mo-Yuen Chow chow@ncsu.edu

Department of Electrical and Computer Engineering

North Carolina State University

Raleigh, NC 27695 USA

This chapter is submitted to IEEE Transactions on Power Systems and is under review.

A NOVEL SAMPLING STRATEGY FOR DISTRIBUTION FAULT DIAGNOSIS: SMALL WORLD STRATIFICATION

Abstract— As an effective tool to expedite the repair process after a power outage, automated distribution fault diagnosis infers the root cause of a fault event by learning from historical faults. Intuitively, only fault events relevant to those under study should be used in the learning process. From the spatial perspective, limiting the study within a small geographic region is preferred in order to focus on the local fault characteristics. However, a small region may not provide sufficient historical fault events for an algorithm to make proper inference.

To cope with this problem, we propose Small World Stratification (SWS) sampling strategy. SWS involves sampling relevant fault events by Geographic Aggregation (GA) and Feature Space Clustering (FSC), and then identifying the group of fault events that should be investigated together. In this paper, we will explain the concept and use simulated fault events to demonstrate the effectiveness of SWS. Experiments show that SWS is necessary to improve the fault diagnosis performance when we focus on a small local region and FSC is superior to GA when fault characteristics in neighboring regions are different.

Index Terms-- discrete event simulation, fault diagnosis, power distribution faults, power system simulation, sampling methods, small-world.

I. INTRODUCTION

Power systems are the vital life lines of modern society, integral to maintaining daily activity and production. As electric devices are used almost everywhere, the cost of interruptions to the power supply is increasingly significant. For some industries, the costs can be as high as several million dollars per hour [1]. Accordingly, a fast service restoration of the power supply is highly desirable to both customers and utility companies. Since the faults in distribution systems account for the majority of customer reliability problems [2], many studies have been devoted to distribution faults. One technology drawing a lot of attention is automated fault diagnosis, which infers the root cause of a fault event based on fault properties. By providing distribution engineers clues before they go on-site to confirm the root cause, automated fault diagnosis is an effective tool to expedite the repairing process.

Automated fault diagnosis learns the relationship between fault properties and its root cause using historical fault events. For example, Zhang *et al.* [3] and Hor *et al.* [4] proposed to use Rough Set theory to discover and represent the relationship between faults and the sequence of protective actions; Nunez *et al.*[5] used features extracted from the fault current and voltage waveforms and environmental factors to diagnose root cause of distribution faults on overhead lines; Xu and Chow *et al.* formulated fault diagnosis as a classification problem and applied several biologically inspired algorithms, including artificial neural networks (ANN), fuzzy systems and artificial immune recognition systems (AIRS) [6-8].

Most of the time, instead of digging into the entire outage database, historical fault records are sampled to investigate a certain fault event. The sampling involves a time window, such as the past a couple of years or certain months prior to the fault event. Meanwhile, the

sampling limits the fault events within a spatial extent, which could be as small as the service region of a distribution lateral or as big as the distribution systems of a major city and neighboring towns.

For electrical characteristics of faults, the sampling process does not affect the study much. For example, the waveform characteristics of tree contact as summarized in [5, 9] are applicable to any distribution systems with similar types of conductors and working conditions, no matter if the lines are from an experimental system in Texas or an actual system serving the Northeastern United States. Moreover, these characteristics can be used for many years before the power line degrades under normal conditions to a point the model is no longer valid.

However, sampling process affects the study of faults when the fault-environment relationships are of the concern. More specifically, diagnosis of the root cause of a fault event needs recent information about the local environments. Fault events that are too old or too far away from the system under investigation generally provide little useful information. In different local systems, even the environmental features needed to diagnose the fault root cause could be different [10]. Thus, proper sampling of historical fault events is crucial in automated fault diagnosis.

In this paper, we will investigate the sampling of historical fault events for distribution fault diagnosis and focus on the spatial aspect. The rest of the paper is structured as follows: Section II formulates the spatial sampling process, defines the resolution of spatial sampling, and discusses the potential data insufficiency problem brought by high spatial resolution; Section III proposes the Small World Stratification (SWS) sampling strategy; Section IV

demonstrates the effectiveness and necessity of SWS using simulated fault events; Section V is the conclusion.

II. SPATIAL SAMPLING IN FAULT DIAGNOSIS

Suppose a fault event e_i occurred at location (x_i, y_i) was caused by root cause c_i. m properties of e_i, $\mathbf{f}_i = [f_{i,1}, f_{i,2}, \ldots, f_{i,m}]$ were identified as related to the root cause. The fault diagnosis process can be described as: given historical fault events with known root causes (*training set*), find a mapping FD: $\mathbf{f} \rightarrow c$ that can map any practical value of $\mathbf{f}$ to the true value of c.

We define a *unit region* to be the smallest geographic area within which fault events are investigated as a group. A unit region ur_j can be the service territory of a feeder or a substation, such as ur_1, ur_2 and ur_3 shown in Fig. 1. Unit regions can be arbitrary polygons on the map with a fixed size as well. ur_4 through ur_7 in Fig. 1 give an example of rectangle grids. Note that unit regions do not necessarily cover the entire area under study and cannot overlap with each other.

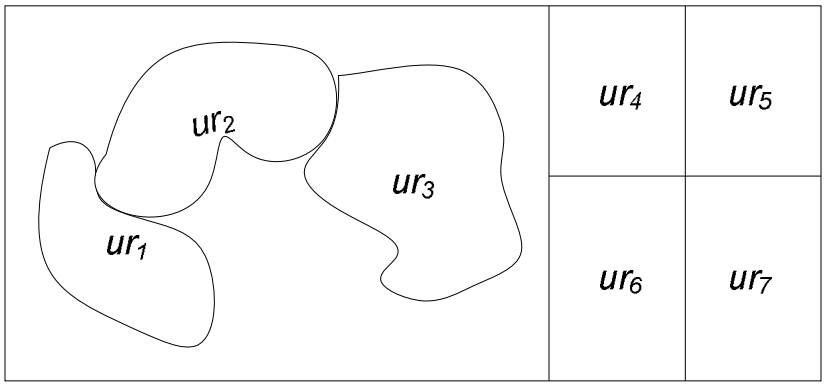


Fig. 1. Examples of unit regions.

In this paper, spatial sampling is the process of extracting a subset of the historical fault events according to the geographic locations. In another word, spatial sampling is to find all the fault events occurred within a certain unit region. A fault event set s_j is defined as:

$$s_j = \{e_i \mid (x_i, y_i) \in ur_j\}. \quad (1)$$

The *resolution of spatial sampling* is defined to be the radius of the smallest circle that covers the smallest unit region. For example, a unit region in Fig. 2 is the service territory of a distribution operation center and the resolution is up to 50 miles.

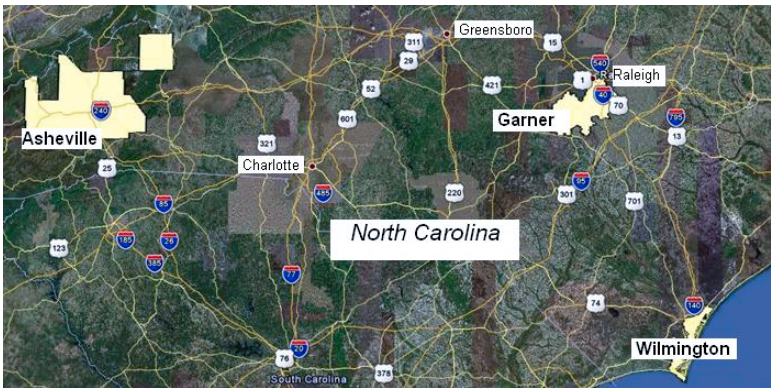


Fig. 2. Unit regions used in Case Study 1 in [10].

Intuitively, high spatial resolution is desired in fault diagnosis in order to capture the local characteristics of the environments and faults. A unit region should be small enough that its fault events can be considered relevant to each other and provide useful information in fault diagnosis. When the spatial resolution is low, fault events occurred in different parts of a unit region could show quite different characteristics. Fig. 3 gives an example of this situation.

Recall that the resolution of spatial sampling shown in Fig. 2 is up to 50 miles. When we zoom into the Garner operation center and use the service territory of a substation as a unit region instead of the service territory of the entire operation center, the spatial resolution

increases to couples of miles. We represent the fault characteristics of a unit region by the Normalized Regional Feature Vector (NRFV) as proposed in [11]. In this case, the NRFV is a two-dimensional vector consisting of the percentage of tree-caused faults and animal-caused faults. Historical fault events during a 5-year period are used to calculate the NRFV and the results are plotted in Fig. 3. The *x*-axis of Fig. 3 is the percentage of tree-caused faults, the *y*-axis represents the percentage of animal-caused faults, and each point represents a unit region. It is obvious that the percentage of tree-caused and animal-caused faults within substations could vary by more than 300% even when they belong to the same operation center.

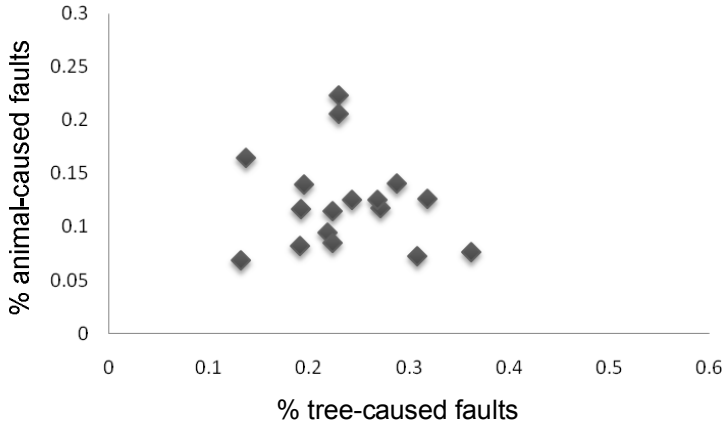


Fig. 3 Fault characteristics within a large unit region could be very different.

However, the resolution of spatial sampling cannot be arbitrarily high considering two factors: spatial resolution of the environmental data and the number of available historical fault events. First, the resolution of spatial sampling should be lower than that of the environmental data. In other words, a unit region should at least cover an area in which environmental factors are distinguishable. Otherwise, faults that occurred within the unit region would show mostly the same properties no matter the root cause. This would make

fault diagnosis practically impossible. With the advancement of information technology and Geographic Information Systems, environmental data with high spatial resolution are becoming available to the general public at a negligible cost (e.g. Google Earth and weather.com). Thus, the number of historical fault events becomes more important in spatial sampling. A very high spatial resolution, i.e. less than half a mile, would provide fairly consistent fault characteristics within the unit region. However, such a unit region only contains a limited number of distribution system components (e.g. poles and transformers). As a result, there are few fault events over the course of a year, given an average service availability of modern distribution systems close to 99.99% [12]. With insufficient historical fault events, it is difficult for an algorithm to properly infer the root cause.

III. SMALL WORLD STRATIFICATION

The concept of *small world* comes from a phenomenon observed in social science. Experiments conducted by Milgram in the 1960's [13, 14] suggest that human society is a small world with a very short average path length – we can find an arbitrary person through on average six mutual acquaintances, which is frequently referred to as "six degrees of separation". Another defining feature of a small world is the large clustering coefficient [15], which means that nodes in a small world tend to form local clusters with connections locally dense but globally sparse.

These two features of small world inspire the idea of Small World Stratification. Suppose the fault event set s_i under investigation is the *target set*. As discussed above, the necessity of small unit regions may lead to so few fault events in s_i that fault diagnosis algorithms do not

work properly. To collect sufficient historical fault events, a *support set* of s_i, s_i^{SP}, is established by connecting fault event sets that are relevant and then identifying the closely connected clusters.

As the first step of SWS, there are two ways to find relevant fault event sets: Geographic Aggregation (GA) and Feature Space Clustering (FSC).

Geographic Aggregation is based on the spatial proximity of unit regions. Assume two unit regions ur_i and ur_j are geographic neighbors as long as they share at least one piece of common boundary, denoted by $ur_i \sim ur_j$. For example, among the unit regions shown in Fig. 1, $ur_5 \sim ur_4$ and $ur_5 \sim ur_7$ as they share boundaries. But ur_5 is not a geographic neighbor of ur_6 because they have only one point in common. Correspondingly, the fault event sets s_i and s_j are called geographic neighbors if the unit regions ur_i and ur_j are geographic neighbors, represented by:

$$ag(s_i, s_j) = \begin{cases} 1 & if\ ur_i \sim ur_j \\ 0 & otherwise \end{cases}. \tag{2}$$

The support set of s_i established by GA is the union of all its geographic neighbors, defined as:

$$s_i^{GA} = \bigcup s_j, \forall ag(s_i, s_j) = 1. \tag{3}$$

GA is usually effective because environmental factors close to each other tend to be similar [16], thus the fault events are more likely to be relevant. Since the geographic locations of fault events are usually recorded in the OMS database, GA is easily applicable to real-world datasets. However, distribution systems may not be similar in neighboring areas. For instance, when one unit region containing large industrial plants is next to unit regions

covering mainly residential areas, GA may not be a good idea. Another limitation of GA is the number of relevant fault events that can be sampled. If more events are needed beyond what the direct neighbors can provide, fault events from neighbors of neighbors should be sampled with caution as they are more likely to have different fault characteristics and be irrelevant to the target set.

Feature Space Clustering considers the fault characteristics within a unit region. Suppose the fault characteristics of the fault event set s_i can be represented by a feature vector $\mathbf{F}_i$, and the difference between fault characteristics can be measured by a distance function $D(\mathbf{F}_i, \mathbf{F}_j)$. The relevance in the feature space is then defined to be:

$$af(s_i, s_j) = \begin{cases} 1 & D(\mathbf{F}_i, \mathbf{F}_j) < d_0 \\ 0 & D(\mathbf{F}_i, \mathbf{F}_j) \geq d_0 \end{cases}, \tag{4}$$

where d_0 is a predefined threshold. The support set of s_i established by FSC is defined to be the union of relevant fault event sets, no matter how far away the corresponding unit regions are:

$$s_i^{FSC} = \bigcup s_j, \forall af(s_i, s_j) = 1. \tag{5}$$

Fig. 4 gives an example of distant unit regions sharing similar fault characteristics. Substations of Asheville operation center are on average 300 to 400 miles away from those of Garner operation center so the fault characteristics of Asheville substations are mostly different from Garner substations. However, the actual historical fault events show that two of them share almost identical characteristics as highlighted by the circle in Fig. 4. By FSC, fault events from these two substations will be identified as relevant and can be used as the support set for each other.

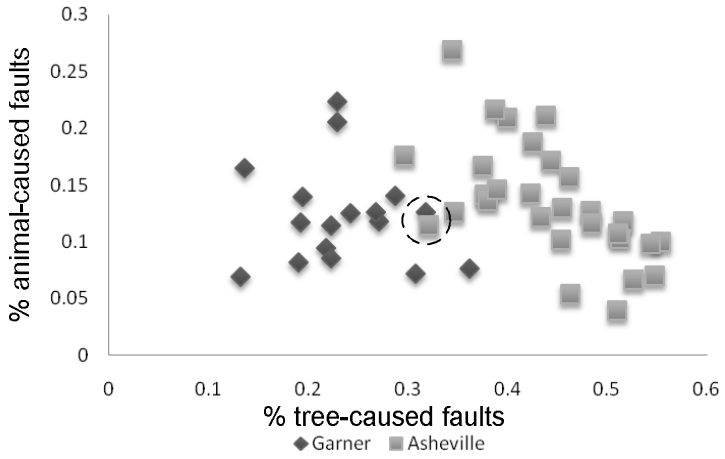


Fig. 4 Fault characteristics could be similar among distant regions.

The idea of SWS is further illustrated in Fig. 5. Areas *A*, *B* and *C* are geographically distant and the unit regions within each area are geographic neighbors. By GA, unit regions in areas *A* and *C* could obtain a support set with fault events from additional two unit regions and this number is three for area *B*. When mapping into the feature space, we notice that all fault event sets in area *A* are relevant to each other and one fault event set from area *B* is similar enough that they form a cluster. By FSC, this fault event set from area *B* could have a support set including all fault events from area *A*. Moreover, strong connections among fault event sets from areas *B* and *C* are observed in the feature space, which suggests they could form a cluster. So fault event sets in *B* and *C* could be the support set for each other, which would potentially double the number of fault events available for fault diagnosis to both areas.

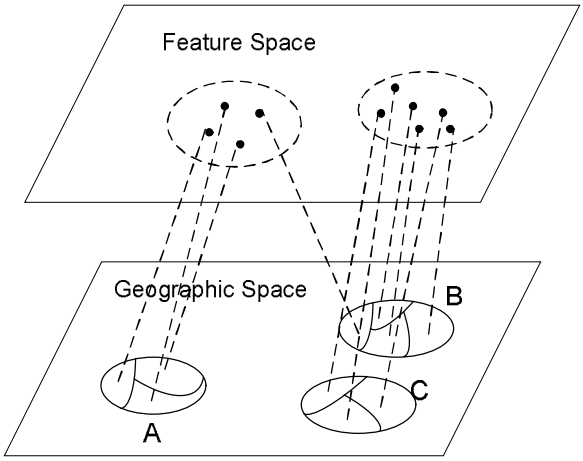


Fig. 5 Small world stratification for fault diagnosis.

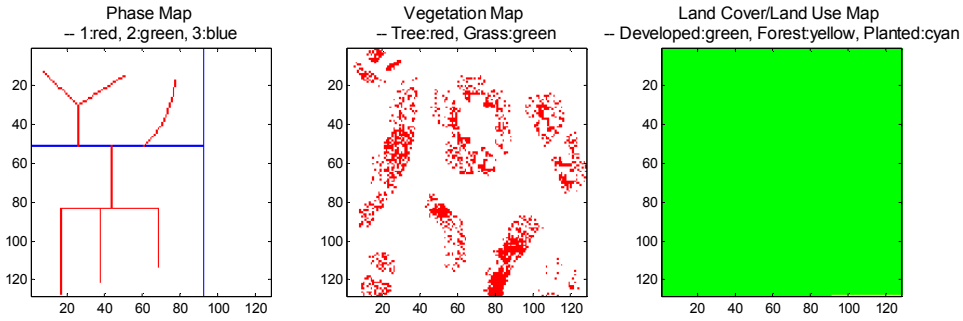
IV. Demonstration of Small World Stratification

A. Simulate Fault Events in Typical Distribution Systems

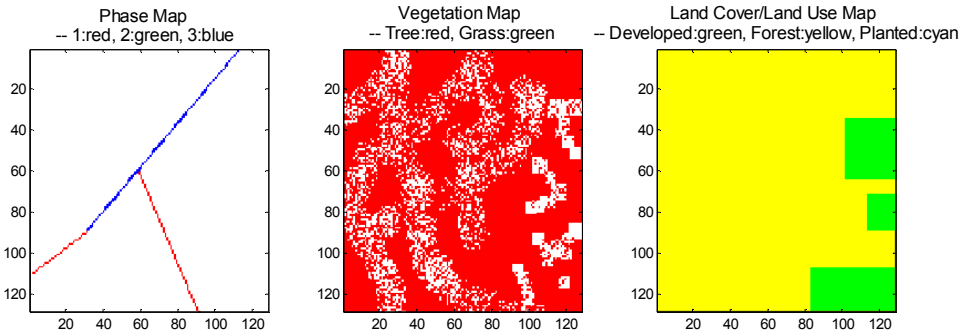
Distribution Fault Simulator is a tool developed to generate realistic fault events given environmental factors and cause-effect relations between environments and faults [17]. The spatial information of the environmental factors is modeled as raster maps. The temporal information, mainly weather conditions, is modeled using probability functions. The causal relation between environmental factors and faults with various root causes are expressed in the form of fuzzy rules. A hierarchical fuzzy inference system is built to infer the probability of tree-caused and animal-caused faults given simulated environments. Fault events are generated and recorded accordingly.

In this paper, we model distribution systems as two typical settings: *metro* and *rural*. A *metro* system is a general model of distribution system in cities and major towns, which is characterized by dense power lines, scattered trees and most of the lands in developed status.

Fault event sets generated from *metro* systems are denoted by *M*. In contrast, a *rural* system has sparse power lines, more trees and most of land as forest. Fault event sets generated from *rural* systems are denoted by *R*. Fig. 6 shows the spatial information of these two typical models. To simplify the demonstration, weather conditions of both models are assumed to be the same, as shown in Fig. 7. An actual service area, such as the service territory of an operation center, can then be modeled as a combination of several *metro* unit regions and *rural* unit regions.



i) *metro* systems.



ii) *rural* systems.

Fig. 6 Spatial environmental information of two typical distribution system models.

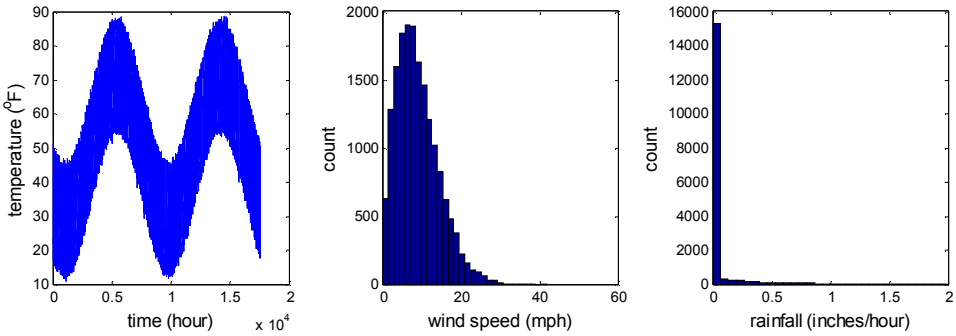


Fig. 7 Weather conditions for the simulation.

B. Experiment Settings

Twelve fault properties generated by simulation are used for fault diagnosis, including:

PHASE (number of phases affected) = {1,2,3};

SEASON (season) = {spring, summer, fall, winter}, each season includes three months and spring starts at March 1st;

TIME (time of the day) = {morning, afternoon, evening, night}, each period includes six hours and morning starts at 6:00am;

DEVICE (protective device activated) = {break/switch, fuse, recloser, sectionalizer, service transformer, source};

WEATHER (weather conditions when the fault occurred) = {clear weather, extreme temperature, raining, snow/ice, thunderstorm, windy};

OH_UG (overhead or underground device) = {OH, UG}.

LULC (land use and land cover type) = {developed, forested upland, shrubland, non-natural woody, herbaceous upland, planted/cultivated, wetlands};

DIST_TREE (distance to the nearest trees);

DIST_ROAD (distance to the nearest major roads);

TEMP (daily maximum temperature);

WIND (hourly average wind speed);

PRECIP (hourly precipitation).

The first seven properties are categorical, which are converted to binary coded dummy variables to represent different levels. Logistic Regression (LR) models and Artificial Neural Networks (ANN) with 30 input variables are used to identify tree-caused faults from non-tree faults, and animal-caused faults from non-animal faults.

The performance of the fault diagnosis is evaluated using Receiver Operating Characteristic (ROC) curves considering the imbalance of data [18]. The area under the ROC curve (AUC) is adopted in this paper to describe the overall performance of the algorithm. It is proven that the value of AUC equals to the probability that an algorithm will report a higher probability of being positive for a randomly chosen positive sample than a randomly chosen negative sample [19]. The ideal value of AUC is 1, which means all the fault events could be diagnosed correctly. AUC value 0.5 indicates that the algorithm could not tell the difference among faults.

To evaluate the fault diagnosis performance, half of the fault events from the target set are randomly selected as the *testing set*, of which the root causes are assumed to be unknown. The remaining half of the target set is combined with its support set as the training set. LR model and ANN network are trained on the training set and then used to diagnose fault events in both the training and testing sets. For each target set, the process is repeated for 30 times and the average AUC is recorded.

C. Test Case 1: Growing Support Sets

With the first test case, we want to demonstrate the effectiveness and necessity of SWS. In other words, we try to find out whether sampling additional fault events from the support set is helpful for fault diagnosis and if it is necessary. Note that we have fault event sets from two different types of distribution systems, *M* and *R*. The following four situations are tested:

a) $s_i \subset M, s_i^{SP} \subset M, i = 1, 2, \cdots, 20$;

b) $s_i \subset M, s_i^{SP} \subset R, i = 1, 2, \cdots, 20$;

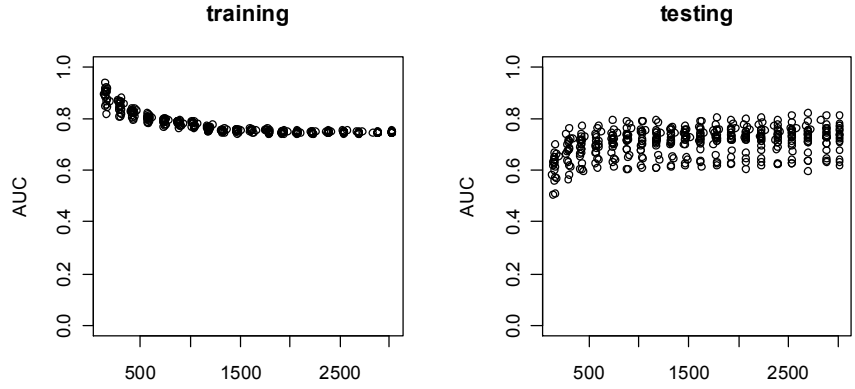
c) $s_i \subset R, s_i^{SP} \subset R, i = 1, 2, \cdots, 20$;

d) $s_i \subset R, s_i^{SP} \subset M, i = 1, 2, \cdots, 20$.

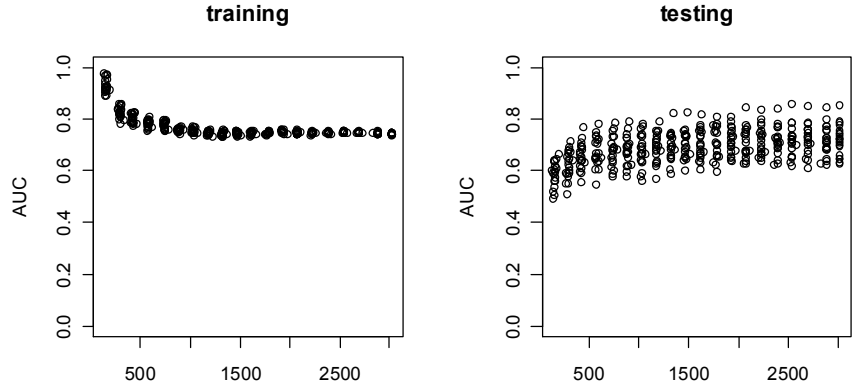
For each target set, support sets with an increasing number of fault events are used. Fig. 8 shows the result of situation a) with LR as an example. The *x*-axis of Fig. 8 represents the number of fault events used in the fault diagnosis and the *y*-axis is the average AUC. Results of the other three situations show similar trends. In general, we observed that:

- The training AUC on the target set is close to 0.9, which suggests very good performance. As more fault events from the support set are used in training, the AUC degrades to around 0.75 and stays at this level. The variation of training AUC reduces as bigger support sets are used.
- The testing AUC when training on the target set only is around 0.6, which is not much better than the random guess. With more fault events in the support set, the testing AUC generally improves in all cases although the trends may not be obvious. The variation of the testing AUC remains at a similar level, if not increases, as

support sets grow bigger. The variation of testing AUC is much larger than that of training.



i) tree-caused faults

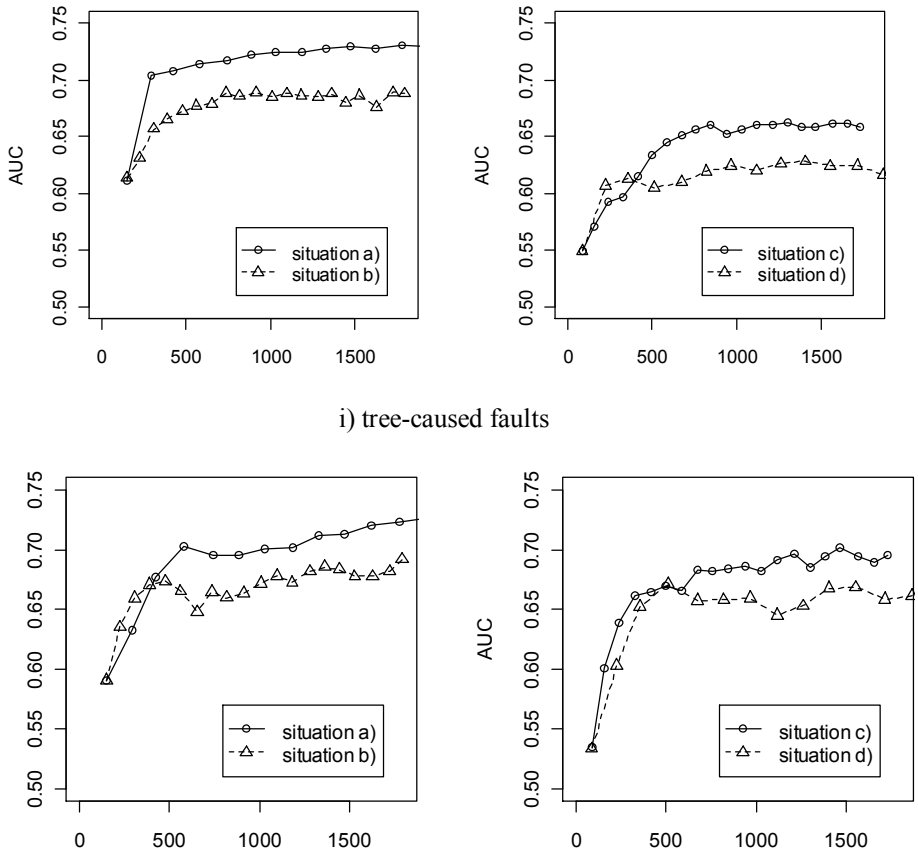


ii) animal-caused faults

Fig. 8. Results with LR of Test Case 1, situation a).

These observations indicate that there are too few fault events in the target set only so the LR model and ANN are over-fitted if trained only on the target set, which leads to poor testing performance. Thus, additional fault events are necessary to make a proper fault diagnosis in a small unit region and SWS strategy effectively improves the performance. To better observe the trend of testing performance, the AUC of 20 target sets for testing under

each situation is averaged by the number of fault events used, and compared side by side as shown in Fig. 9 and 10.



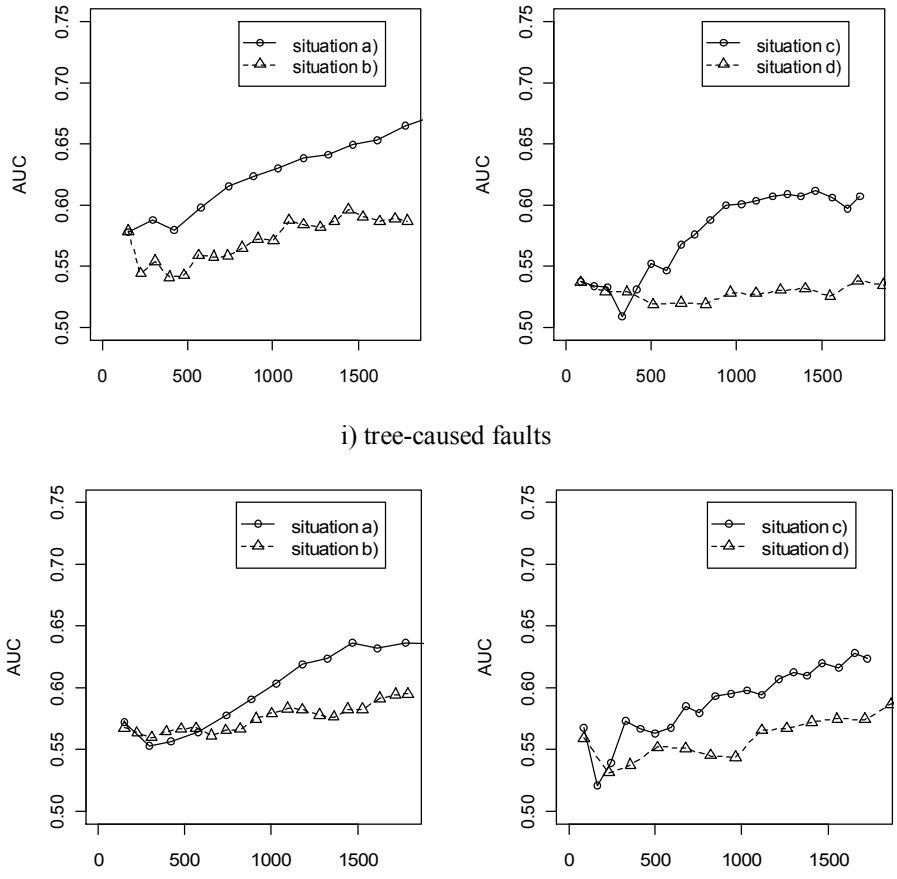
i) tree-caused faults

ii) animal-caused faults

Fig. 9. Comparison of average testing performance of LR.

As we can see in Fig. 9, it is quite clear that sampling additional fault events for training improves the testing performance under all situations. A support set with fault events from similar unit regions (situations *a* and *c*) provides greater improvement compared to the support sets including fault events from different unit regions (situations *b* and *d*). The results with ANN (Fig. 10) show similar trends for sampling from similar fault event sets while sampling from different fault event sets provides little help. In general, we can conclude that

SWS is effective in improving fault diagnosis performance while sampling from similar fault event sets (FSC) is more efficient.



i) tree-caused faults

ii) animal-caused faults

Fig. 10. Comparison of average testing performance of ANN.

D. Test Case 2: GA vs. FSC

With Test Case 2, we evaluate GA and FSC to find out how these two differ in improving fault diagnosis performance. Two hypothetical service areas are used in this test case. As shown in Fig. 11, service area *A* consists of 6 *metro* unit regions and 6 *rural* unit regions and service area *B* contains 3 *metro* unit regions and 8 *rural* regions. Both service areas model a distribution system that serves an urban area and its surroundings of roughly the same size

while the urban areas in A take more space than in B. Power lines in both service areas are shown in Fig. 12.

Service area A is assumed to be far away from B so GA is applicable only within each service area. The support set built by GA is defined in (3). Since fault events from the same type of unit regions are generated from the same simulation settings, we know that those fault event sets $s_i \subset M, i = 1,2,\cdots,9$ are similar to each other and $s_j \subset R, j = 1,2,\cdots,14$ are similar to each other. By FSC, the support set of a target set s_i is:

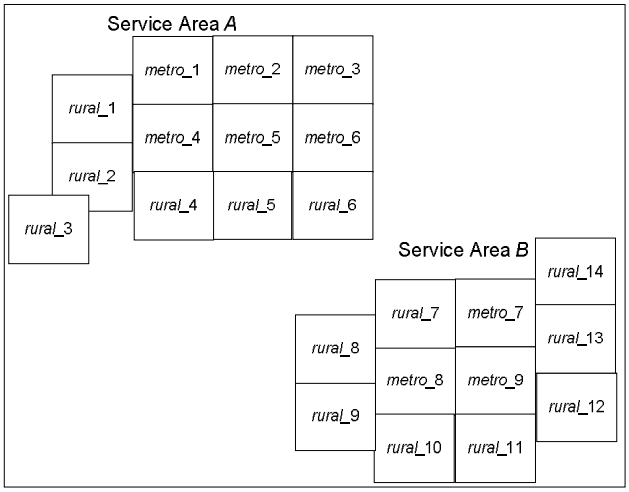
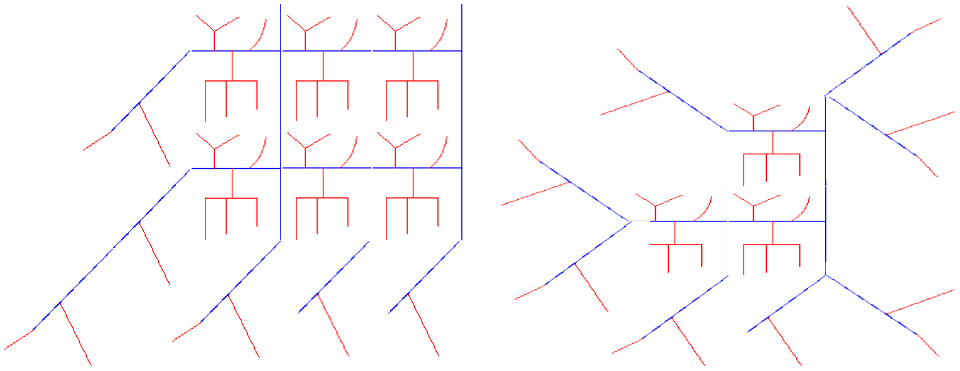


Fig. 11. Two hypothetical service areas consisting of *metro* and *rural* unit regions.



i) Service area A ii) Service area B

Fig. 12. Power lines in the hypothetical service areas.

$$s_i^{FSC} = \begin{cases} \bigcup s_t, s_t \subset M \text{ and } t \neq i & \text{if } s_i \subset M \\ \bigcup s_t, s_t \subset R \text{ and } t \neq i & \text{if } s_i \subset R \end{cases}. \quad (6)$$

Typical fault event sets with sufficient geographic neighbors from both service areas are selected for comparison. Specifically, *M4* from *A* has 5 neighbors, *R2* from *A* has 4 neighbors, *M9* from *B* has 5 neighbors and *R13* from *B* has 4 neighbors.

Though AUC represents the overall performance of an algorithm, the fault diagnosis decision must be made according to a decision threshold which corresponds to one point on the ROC curve. In this test case, the G-mean measure previously used in [6-8] is adapted. With the fault diagnosis decision made, every fault event in the testing set will fall into one of the four situations listed in Table I.

TABLE I

CONFUSION MATRIX

	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	True Positive (TP)	False Negative (FN)
Actual Negative Class	False Positive (FP)	True Negative (TN)

By counting the number of fault events in each situation, the G-mean measure is calculated as:

$$G-mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}. \quad (7)$$

The decision threshold is selected to maximize the G-means on training sets. The average testing G-means of diagnosing faults based only on the target set, sampling by GA and sampling by FSC are summarized in Table II and III with standard deviation in the brackets.

They are compared by two-sample *t*-test at the 0.05 significance level, which roughly indicates that the conclusion about the performance difference is correct 95% of all the time.

TABLE II

FAULT DIAGNOSIS PERFORMANCE OF LR MEASURED BY G-MEAN

Target Set		Target Set Only (TO)	GA	FSC	Comparison by t-test
M4	T	0.614(0.058)	0.614(0.077)	0.694(0.042)	TO~GA<FSC
R2		0.521(0.074)	0.594(0.060)	0.614(0.065)	TO<GA~FSC
M9		0.558(0.100)	0.703(0.058)	0.640(0.060)	TO<FSC<GA
R13		0.486(0.051)	0.482(0.060)	0.616(0.062)	TO~GA<FSC
M4	A	0.507(0.141)	0.733(0.077)	0.736(0.053)	TO<GA~FSC
R2		0.465(0.116)	0.575(0.077)	0.665(0.085)	TO<GA<FSC
M9		0.376(0.195)	0.653(0.087)	0.632(0.100)	TO<GA~FSC
R13		0.383(0.273)	0.746(0.038)	0.654(0.157)	TO<FSC<GA

T: tree-caused faults, A: animal-caused faults.

TABLE III

FAULT DIAGNOSIS PERFORMANCE OF ANN MEASURED BY G-MEAN

Target Set		Target Set Only (TO)	GA	FSC	Comparison by t-test
M4	T	0.498(0.068)	0.555(0.098)	0.644(0.063)	TO<GA<FSC
R2		0.475(0.089)	0.507(0.061)	0.615(0.065)	TO~GA<FSC
M9		0.591(0.096)	0.619(0.096)	0.627(0.062)	TO~GA~FSC
R13		0.465(0.060)	0.472(0.086)	0.553(0.064)	TO~GA<FSC
M4	A	0.423(0.155)	0.696(0.066)	0.719(0.075)	TO<GA~FSC
R2		0.513(0.165)	0.450(0.185)	0.532(0.134)	TO~GA~FSC
M9		0.172(0.221)	0.604(0.074)	0.593(0.092)	TO<GA~FSC
R13		0.364(0.299)	0.545(0.232)	0.689(0.109)	TO<GA<FSC

T: tree-caused faults, A: animal-caused faults.

From Table II and III we observed that:

- GA effectively improves the fault diagnosis performance with LR most of the time, except for diagnosing tree-caused faults in M4 and R13. With ANN, GA improves the fault performance significantly only half of the time. As we found out in Test Case 1, the neighboring unit regions of a different type affect GA.

- FSC improves the fault diagnosis performance significantly in all cases, except for diagnosing tree-caused faults in M9 and animal-caused faults in R2 by ANN. With ANN, it provides comparable or greater improvement than GA. This is consistent with what we found out in Test Case 1 -- sampling from similar fault event sets provides better support to local fault diagnosis.

V. CONCLUSION

In distribution fault diagnosis, small unit regions are preferred to study the local fault characteristics. However, with small unit regions, the historical fault events may not be sufficient for an algorithm to make proper inference. To cope with this problem, we propose Small World Stratification to sample additional fault events from other unit regions while keeping every unit region small.

The idea of SWS involves building a support set for the fault event set under study by finding relevant fault event sets with Geographic Aggregation and Feature Space Clustering, and identifying the closely connected cluster. This sampling strategy is demonstrated with simulated fault events. Experimental results show that using SWS to collect sufficient fault events is necessary to improve the fault diagnosis performance and FSC could provide greater improvements than GA when fault characteristics in neighboring unit regions are different.

REFERENCES

[1] Electricity Advisory Committee, US Department of Energy, "Smart grid: enabler of the new energy economy," [Online]. Available: http://www.oe.energy.gov/DocumentsandMedia/final-smart-grid-report.pdf.

[2] R. E. Brown, *Electric Power Distribution Reliability*. New York: Marcel Dekker, Inc, 2002.

[3] Q. Zhang, Z. Han, and F. Wen, "A new approach for fault diagnosis in power systems based on rough set theory," presented at the 4th Int. Conf. Advances in Power System Control, Operation and Management, Hong Kong, China, 1997.

[4] C.-L. Hor, P. A. Crossley, and S. J. Watson, "Building Knowledge for Substation-Based Decision Support Using Rough Sets," *IEEE Trans. Power Delivery,* vol. 22, pp. 1372-1379, 2007.

[5] V. B. Nunez, S. Kulkarni, S. Santoso, and J. Melendez, "Feature analysis and classification methodology for overhead distribution fault events," in *IEEE Power and Energy Society General Meeting 2010,* Minneapolis, MN, 2010.

[6] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Trans. Power Systems,* vol. 21, pp. 53-60, 2006.

[7] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm," *IEEE Trans. Power Systems,* vol. 22, pp. 164-171, 2007.

[8] L. Xu, M.-Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using Artificial Immune Recognition System (AIRS) algorithm," *IEEE Trans. Power Systems,* vol. 22, pp. 198-204, 2007.

[9] J. A. Wischkaemper, C. L. Benner, and B. D. Russell, "Electrical characterization of vegetation contacts with distribution conductors - investigation of progressive fault behavior," in *IEEE/PES Transmission and Distribution Conference and Exposition*, Chicago, IL, 2008, pp. 1-8.

[10] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Statistical feature selection from massive data in distribution fault diagnosis," *IEEE Trans. Power Systems,* vol. 25, pp. 642-648, 2010.

[11] Y. Cai and M.-Y. Chow, "Similarity measures in small world stratification for distribution fault diagnosis," submitted to *IEEE Trans. Power Systems*.

[12] J. Zhong, C. Wang, and Y. Wang, "Chinese growing pains," *IEEE Power and Energy Magazine,* vol. 5, pp. 33-40, 2007.

[13] S. Milgram, "The small world problem," *Psychology Today,* vol. 2, pp. 60-67, 1967.

[14] J. Travers and S. Milgram, "An experimental study of the small world problem," *Sociometry,* vol. 32, pp. 425-443, 1969.

[15] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature,* vol. 393, pp. 440-442, 1998.

[16] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, *Geographic Information Systems and Science*. Chichester, England: Wiley, 2001.

[17] Y. Cai and M.-Y. Chow, "Cause-effect modeling and spatial-temporal simulation of power distribution fault events," *IEEE Trans. Power Systems,* accepted.

[18] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Evaluation of distribution fault diagnosis algorithms using ROC curves," in *IEEE Power and Energy Society General Meeting 2010,* Minneapolis, MN, 2010.

[19] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters,* vol. 27, pp. 861-874, 2006.

CHAPTER VI

MEASURING SIMILARITY AMONG REGIONS FOR DISTRIBUTION FAULT DIAGNOSIS

Yixin Cai Mo-Yuen Chow

ycai2@ncsu.edu chow@ncsu.edu

Department of Electrical and Computer Engineering

North Carolina State University

Raleigh, NC 27695 USA

This chapter is submitted to IEEE Transactions on Power Systems and is under review.

MEASURING SIMILARITY AMONG REGIONS FOR DISTRIBUTION FAULT DIAGNOSIS

Abstract— Distribution fault diagnosis algorithms learn from historical fault events and infer the root cause of the fault under study in order to expedite the service restoration after a power outage. To cope with the problem of insufficient historical data when diagnosing faults in a small local area, we have proposed a sampling strategy, Small World Stratification (SWS). SWS involves sampling relevant fault events by Geographic Aggregation (GA) and Feature Space Clustering (FSC), and identifying the group of fault events that should be investigated together.

FSC identifies relevant fault event sets according to how similar they are, where measuring the similarity among regions is essential. In this paper, we propose four regional feature vectors (RFV): normalized regional feature vectors (NRFV), relative regional feature vectors (RRFV), likelihood regional feature vectors (LRFV) and generalized regional feature vectors (GRFV), derived from measures used to analyze distribution faults. Similarity measures based on the distance between RFVs are evaluated using fault events simulated by the Distribution Fault Simulator. Experimental results suggest that GRFV is the best among the four.

Index Terms— clustering methods, discrete event simulation, fault diagnosis, power distribution faults, power system simulation, sampling methods, small world stratification.

I. INTRODUCTION

Distribution fault diagnosis identifies the root cause of faults in power distribution systems before the affected power lines can be reenergized. To expedite this process, automated fault diagnosis has been studied for decades. Different algorithms are used to identify the root cause of a fault based on the fault properties and environmental factors. For example, Chow and Taylor [1] analyzed tree-caused and animal-caused faults in distribution systems with four different measures to reveal the relationship between the environmental factors and the root cause; Chen *et al.* [2] used a cause-effect network and fuzzy sets to find the root cause based on protective device settings and their operations during the fault in distribution substations; Xu and Chow *et al.* [3-5] formulated fault diagnosis as a classification problem and applied several biologically inspired algorithms, including artificial neural networks (ANN), fuzzy systems and artificial immune recognition systems (AIRS); Nunez *et al.* [6] used features extracted from the fault current and voltage waveforms and environmental factors to diagnose root causes of distribution faults on overhead lines.

In order to infer the root cause, historical faults are needed for the automated fault diagnosis algorithms to learn the relationship between environmental factors and the fault root cause. Considering that distribution systems are spatially dispersed and heterogeneous, investigating fault events within a small local area is preferred to focus on the local fault characteristics. However, with small regions, the historical fault events may not be sufficient for an algorithm to make proper inference.

To cope with this problem, we have proposed Small World Stratification to sample additional fault events from other regions while keeping the study regions small [7]. The idea of SWS involves finding relevant fault events by Geographic Aggregation (GA) and Feature Space Clustering (FSC), and identifying the group of fault events that should be investigated together. GA finds relevant fault event sets based on the geographic proximity of the regions under study, while FSC looks for fault event sets that are close in the feature space. In our previous research demonstrating SWS, we assume that the characteristics of a fault event set are known so that fault events from the same type of regions can be sampled. However, applying FSC to the real-world data is always challenging as the measures of similarity among regions are not readily available.

In this paper, we will explore different ways of measuring the similarity among regions. In Section II, we adapt normalized measures, relative measures, and likelihood measures previously used in analyzing distribution faults to define Regional Feature Vectors (RFV) and measure similarity among unit regions by the distance between RFVs. In Section III, we use fault events generated by the Distribution Fault Simulator [8] to evaluate these similarity measures. Section IV discusses the results and draws conclusions.

II. Measures Of Similarity Among Unit Regions

A. Problem Formulation

Suppose a fault event e_j occurred at location (x_j, y_j). We call the smallest geographic area within which fault events are investigated as a group a *unit region*. Fault events that occur within the unit region ur_i are called a fault event set s_i, defined as:

$$s_i = \{e_j \mid (x_j, y_j) \in ur_i\}. \quad (1)$$

The idea of FSC is to find relevant fault event sets based on how similar their fault characteristics are. Suppose the fault characteristics of the fault event set s_i can be represented by a feature vector $\mathbf{F}_i$, and the difference between fault characteristics can be measured by a distance function $D(\mathbf{F}_i, \mathbf{F}_j)$. The relevance between fault event sets in the feature space is then defined as:

$$af(s_i, s_j) = \begin{cases} 1 & D(\mathbf{F}_i, \mathbf{F}_j) < d_0 \\ 0 & D(\mathbf{F}_i, \mathbf{F}_j) \geq d_0 \end{cases}, \quad (2)$$

where d_0 is a predefined threshold. By using FSC, additional fault events can be sampled from the union of similar fault event sets, no matter how far away the corresponding unit regions are:

$$s_i^{FSC} = \bigcup s_j, \forall af(s_i, s_j) = 1. \quad (3)$$

The problem we discuss in this paper is how to define a proper $\mathbf{F}_i$ for the fault event set s_i (we call it a Regional Feature Vector hereafter as it represents the fault characteristics of a unit region) and the corresponding distance function D.

B. Regional Feature Vectors

Normalized measures, relative measures and likelihood measures of fault events within a study region have been used in analyzing distribution faults [1, 9]. Normalized Regional Feature Vectors (NRFV), Relative Regional Feature Vectors (RRFV) and Likelihood Regional Feature Vectors (LRFV) are derived from these measures.

The normalized measure of fault events with a certain root cause is the fraction of fault events with the root cause of interest over the total number of fault events under study.

Suppose we are interested in m different root causes [c_1, c_2, ..., c_m]. The normalized measure of fault events is defined as:

$$NOR_i = \frac{N(r = c_i)}{N}, \tag{4}$$

where $N(r=c_i)$ is the number of fault events with root cause c_i and N is the total number of fault events under study. Correspondingly, the normalized regional feature vector of a fault event set is defined as:

$$NRFV = \left[NOR_1, NOR_2, \cdots, NOR_m\right]. \tag{5}$$

The relative measure of fault events represents the fraction of fault events with a certain root cause under different specified conditions. Suppose we have k properties [f_1, f_2, ..., f_k] associated with a fault event and the jth property takes l_j discrete values $[a_{j,1}, a_{j,2}, \cdots a_{j,l_j}]$. The relative measure of fault events is defined as:

$$REL_{i,j,t} = \frac{N(r = c_i, f_j = a_{j,t})}{N(r = c_i)}, t \in [1, 2, \cdots l_j], \tag{6}$$

where $N(r=c_i, f_j = a_{j,t})$ is the number of fault events with root cause c_i and the property f_j is equal to the value $a_{j,t}$, and $N(r=c_i)$ is the total number of fault events with root cause c_i. Thus, the relative regional feature vector for fault events with root cause c_i is defined as a vector with $\sum_{j=1}^{k} l_j$ components:

$$RRFV_i = [REL_{i,1,1}, REL_{i,1,2}, \cdots, REL_{i,1,l_1}, REL_{i,2,1}, \cdots REL_{i,2,l_2}, \cdots, REL_{i,k,1} \cdots REL_{i,k,l_k}]. \tag{7}$$

The likelihood measure of fault events is the conditional probability of fault events having root cause c_i given a specific condition, defined as:

$$LIK_{i,j,t} = \frac{N(r = c_i, f_j = a_{j,t})}{N(f_j = a_{j,t})}, t \in [1, 2, \cdots l_j], \quad (8)$$

where $N(f_j = a_{j,t})$ is the number of all fault events when $f_j = a_{j,t}$. Similar to RRFV, the likelihood regional feature vector for faults with root cause c_i has $\sum_{j=1}^{k} l_j$ components, as follows:

$$LRFV_i = [LIK_{i,1,1}, LIK_{i,1,2}, \cdots, LIK_{i,1,l_1}, LIK_{i,2,1}, \cdots LIK_{i,2,l_2}, \cdots, LIK_{i,k,1} \cdots LIK_{i,k,l_k}]. \quad (9)$$

Moreover, a Generalized Regional Feature Vector (GRFV) contains selected components from the previous three RFVs, in the form of

$$GRFV_i = [NOR_{i_1}, NOR_{i_2}, \cdots, REL_{i,j_1,1}, \cdots, REL_{i,j_1,l_{j_1}}, LIK_{i,j_2,1}, \cdots, LIK_{i,j_2,l_{j_2}}]. \quad (10)$$

C. Distance among Unit Regions in the Feature Space

As the characteristics of fault event sets are represented by the RFVs, the difference among them can be measured by the Euclidean distance among RFVs. Specifically, we can measure the difference of two fault event sets *u* and *v* by:

- Distance of NRFV: $DN = \left\| NRFV_u - NRFV_v \right\| = \sqrt{\sum_{i=1}^{m} (NOR_{u,i} - NOR_{v,i})^2}$.
- Distance of RRFV in terms of root cause c_i: $DR_i = \left\| RRFV_{u,i} - RRFV_{v,i} \right\|$.
- Distance of LRFV in terms of root cause c_i: $DL_i = \left\| LRFV_{u,i} - LRFV_{v,i} \right\|$.
- Distance of GRFV in terms of root cause c_i: $DG_i = \left\| GRFV_{u,i} - GRFV_{v,i} \right\|$.

The similarity among unit regions is opposite to these distances: the smaller the distance, the more similar. Thus, we will not distinguish between similarity measure and distance in the feature space in the rest of this paper where it does not cause confusion.

III. Evaluation of Similarity Measures

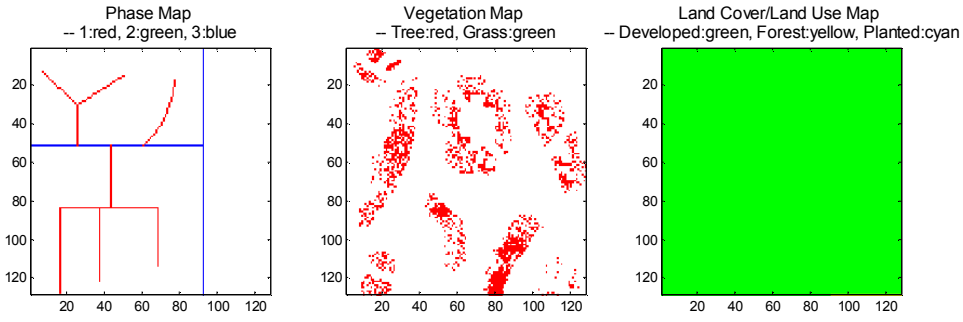
A. Modeling Typical Distribution Systems

To evaluate the similarity measures, the Distribution Fault Simulator [8] is used to generate realistic fault events from different unit regions. With the distribution fault simulator, the spatial information of significant environmental factors leading to faults is modeled as a series of raster maps. The temporal information of weather conditions is modeled as a set of probability functions. The causal relationships among these factors and faults with various root causes are investigated and expressed in the form of fuzzy rules. A hierarchical fuzzy inference system is then built to infer the probability of faults in the simulated environments. Fault events are generated and recorded accordingly.

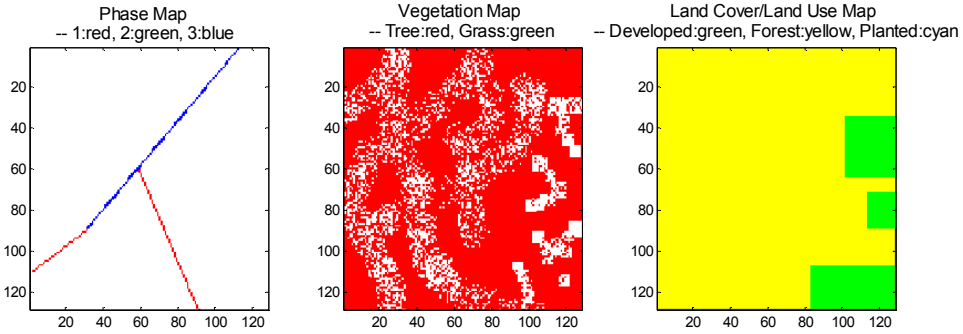
By changing the environmental settings, we are able to generate fault events for different unit regions. The difference among unit regions can be categorized into three aspects: spatial difference, temporal difference and logical difference.

Spatial difference refers to the difference among unit regions in terms of spatial information. In this paper, we model the spatial aspect of distribution systems as two typical types: *metro* and *rural*. A *metro* system is a general model of distribution systems in cities and major towns, which is characterized by dense power lines, scattered trees, with most of

the land *developed*. In contrast, a *rural* system has sparse power lines, more trees and most of land is *forest*. Fig. 1 shows the spatial information of these two typical models.



a) *metro* systems.



b) *rural* systems.

Fig. 1. Spatial environmental information of two distribution system models.

Temporal difference is the difference in weather characteristics under the current implementation of the simulator. Two typical weather conditions are used. The *basic* weather condition is defined to be an average annual temperature 50 ˚F, an average hourly wind speed of 10mph and an average rainfall intensity of 0.55 inches per hour with a probability 0.15 of rainfall. In contrast, the *nice* weather condition is warmer (an average temperature of 65 ˚F), less windy (windy 80% of the time and an hourly average of 5mph) and drier (an average rainfall intensity of 0.15 inches per hour with a rainfall probability of 0.05). Both weather conditions during a two-year simulation period are shown in Fig. 2.

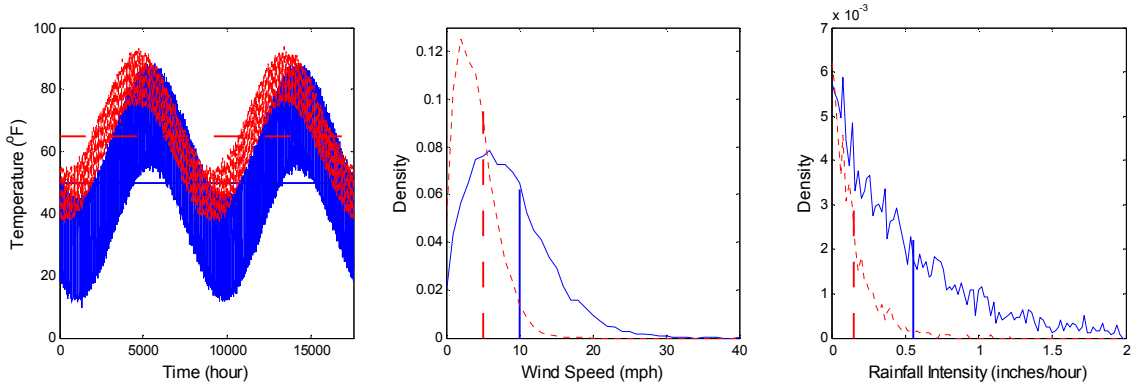
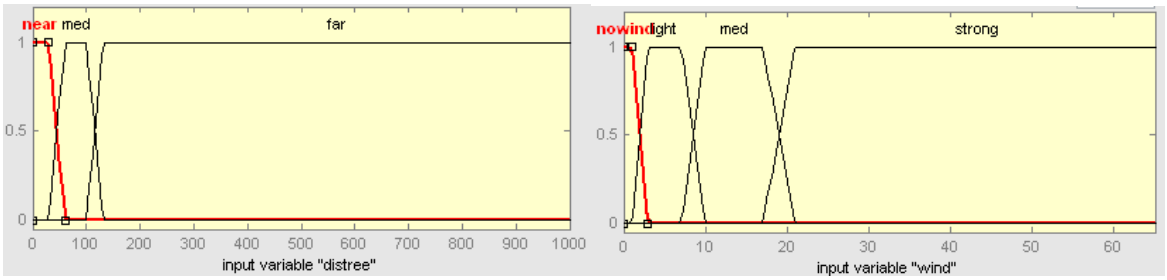
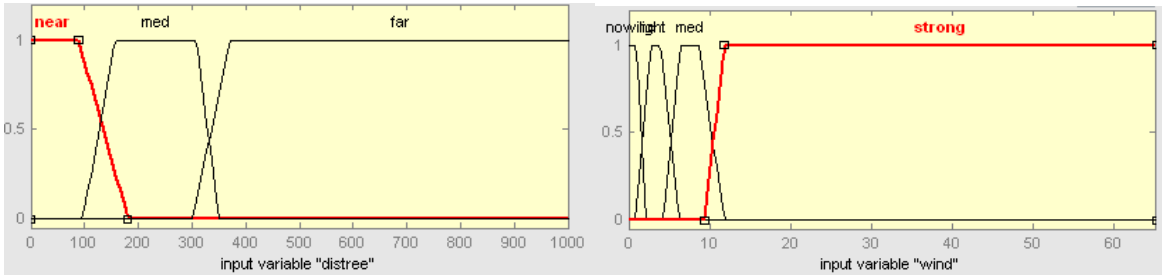


Fig. 2. Two typical weather conditions. (Blue solid lines are *basic* condition and red dotted lines are *nice* condition)

Logical difference refers to the difference in the membership functions representing the environmental factors and the rules that govern the probability of fault events. Usually, rules representing the cause-effect relationships between environmental factors and faults do not change, while membership functions may need to be changed when modeling different environments. For evaluation purpose, two sets of membership functions are defined. The distance to the nearest trees is represented by three fuzzy sets, *near*, *med*, and *far*, and the wind speed is represented by *nowind*, *light*, *med*, and *strong*. Under the *normal* conditions, distances within 30 to 60 meters are defined as *near*, those of more than 100 meters are considered as *far*, and distances in-between are regarded as *med*. The level of wind speed is defined according to the commonly used Beaufort scale [10], as shown in Fig. 3a). Under the *extreme* conditions, the *near* distance encompasses a much larger range, up to more than 100 meters. The *strong* wind covers wind speeds as low as 10mph, as shown in Fig. 3b). Both changes under the *extreme* conditions are in favor of tree-caused faults, which means more tree-caused faults are likely to be generated under the same spatial and temporal settings.



a) *normal* situation.



b) *extreme* situation.

Fig. 3 Two sets of membership functions.

B. Comparison of Similarity Measures by Clustering

In reality, two unit regions can be different in any combination of these three aspects. To evaluate the similarity measures, we defined 6 types of unit regions and changed only one aspect at a time. Unit regions *metro* and *rural* are under *basic* weather conditions, with *normal* membership functions. Unit regions *metro_nice* and *rural_nice* are under *nice* weather conditions with *normal* membership functions, and *metro_extreme* and *rural_extreme* are defined as *extreme* membership functions under *basic* weather conditions. Ten fault event sets of each type were generated.

For this research, we are interested in the faults caused by tree and animals. The NRFV in this case is two dimensional, and consists of the normalized measure of tree-caused and animal-caused faults. The NRFVs of the 60 fault event sets are plotted in Fig. 4, where the *x*-

axis is the normalized measure of tree-caused faults, *y*-axis is the normalized measure of animal-caused faults, and each symbol represents a fault event set.

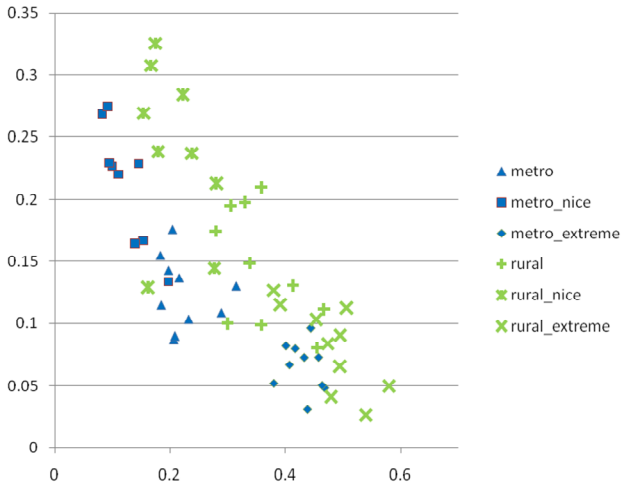


Fig. 4 Six types of fault event sets are used for the evaluation.

Seven categorical fault properties are used to define the RRFV and LRFV, including:

PHASE (number of phases affected) = {1,2,3};

SEASON (season) = {spring, summer, fall, winter};

TIME (time of the day) = {morning, afternoon, evening, night};

DEVICE (protective device activated) = {break/switch, fuse, recloser, sectionalizer, service transformer, source};

WEATHER (weather conditions when the fault occurred) = {clear weather, extreme temperature, raining, snow/ice, thunderstorm, windy};

OH_UG (overhead or underground device) = {OH, UG}.

LULC (land use and land cover type) = {developed, forested upland, shrubland, non-natural woody, herbaceous upland, planted/cultivated, wetlands}.

The RRFV and LRFV in this test both have 32 components. The GRFV tested in this paper is a combination of all available components from NRFV, RRFV and LRFV, and has 66 components in total.

As we can see in Fig. 4, fault event sets of the same type tend to cluster together in the normalized regional feature space though clusters of different types do overlap. Some of the clusters can be identified intuitively by observing the plot. However, the high dimensional feature spaces for the RRFV, LRFV and GRFV are difficult to visualize. Hence, we use clustering methods to compare how these similarity measures differ from each other in finding the correct clusters of fault event sets.

K-means clustering is widely used to divide data into k groups by minimizing the sum of the squared distance from data points to their cluster centers [11]. With the same data and the same algorithm, a good similarity measure would yield clusters that best reflect reality.

The *normalized mutual information* is used to measure how accurate the detected clusters are compared to the actual ones, as suggested by Danon and Diaz-Guilera *et al.* [12]. Suppose there are n_a actual clusters A and n_k detected clusters K. The normalized mutual information is defined as:

$$NMI(A,K) = \frac{-2\sum_{i=1}^{n_a}\sum_{j=1}^{n_k} N_{ij}\log(\frac{N_{ij}N}{N_{i\cdot}N_{\cdot j}})}{\sum_{i=1}^{n_a} N_{i\cdot}\log\frac{N_{i\cdot}}{N} + \sum_{j=1}^{n_k} N_{\cdot j}\log\frac{N_{\cdot j}}{N}}, \tag{11}$$

where N is the total number of fault event sets to be clustered, N_{ij} represents the number of fault event sets that are actually in cluster i and detected to be in cluster j, $N_{i\cdot}$ and $N_{\cdot j}$ are the sum of all N_{ij} where the actual cluster is i and where the detected cluster is j, respectively.

NMI measures the information about the actual clusters provided by the detected clusters, with values ranging from 0 to 1. NMI = 0 indicates that the actual clusters are independent from the detected clusters and the detected clusters do not provide any information regarding the actual ones. NMI = 1 represents the detected clusters completely reflect the actual situation.

The simulated fault events are divided into several subsets. Each of the subsets contains 20 fault event sets of two different types and is different in only one aspect. K-means clustering is applied to every subset to divide it into two groups and the results are summarized in Table I.

TABLE I

ACCURACY OF CLUSTERS DETECTED BY K-MEANS CLUSTERING MEASURED BY NMI

		Similarity Measured by			
		DN	DR	DL	DG
Spatial Difference	*metro* vs. *rural*	0.619	1	0.510	1
	metro_nice vs. *rural_nice*	0.619	1	0.275	1
	metro_extreme vs. *rural_extreme*	0.119	1	0.619	1
Temporal Difference	*metro* vs. *metrol_nice*	0.510	1	1	1
	rural vs. *rural_nice*	0.510	1	0.421	0.619
Logical Difference	*metro* vs. *metro_extreme*	1	0	1	1
	rural vs. rural_extreme	0.192	0.011	0.210	0.275

From table I we can see that DR is a good measure for the difference in spatial and temporal information but does not perform well on logical difference. DL is better than DR in detecting the difference caused by different membership functions. In general, DG is the best of all four similarity measures.

C. Performance of FSC with Different Similarity Measures

In this section, we will test the similarity measures in a more complex setting and investigate how they affect the performance of FSC. All of the simulated fault events are used in this test. Suppose we are interested in fault diagnosis in a *metro* unit region (the *target region*). The similar regions in FSC are identified by K-means clustering given six clusters. Fault event sets in the same cluster as the target region are used as the support set. Note that K-means clustering is not necessarily the tool used for FSC. In practice, it is usually difficult to apply K-means clustering when the actual number of clusters is unknown. Each of the 10 *metro* unit regions is tested and the results are summarized in Fig. 5 and 6.

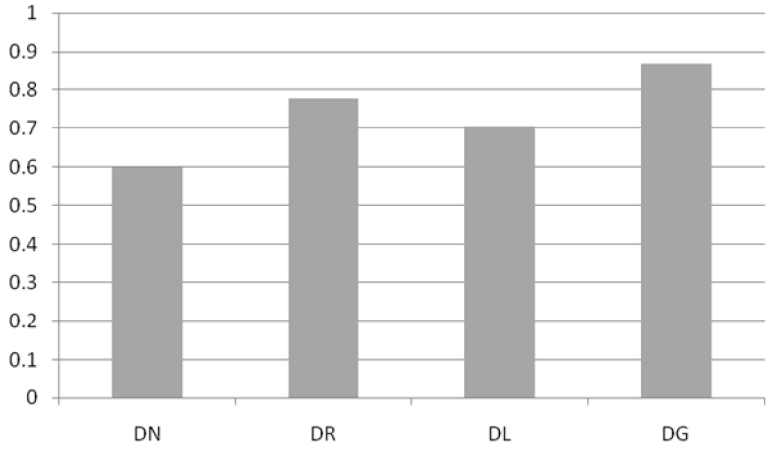


Fig. 5. Accuracy of clusters detected by K-means clustering measured by NMI.

Fig. 5 shows the NMI of clusters found by K-means using different similarity measures. Fig. 6 shows the percentage of fault events from the same type of unit regions as the target region among all the fault events sampled by FSC. We can see that DG performs best with the highest NMI and an average percentage close to 100%, which means similar regions identified by K-means using DG are almost correct except for the region *metro 10*.

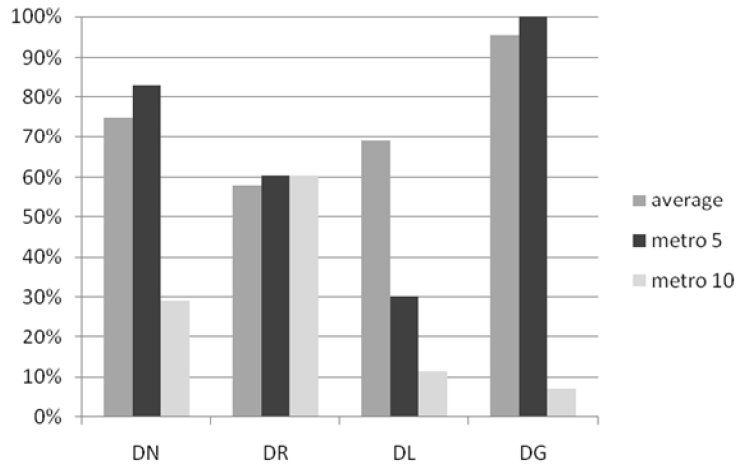


Fig. 6. Percentage of fault events from the same type of regions among all the fault events sampled by FSC.

Fig. 7 shows the performance improvements in terms of the percentage increase of AUC [13]. As a reference, the performance increase of sampling additional fault events from different type of regions (AD) as well as from the same type of regions (AS) are plotted alongside. As we can see, FCS with 4 different similarity measures gives a performance improvement of between 20% and 24%. Although the difference among them is small, we notice that DG leads to better improvement than others, close to the ideal situation (AS). In some cases, DG could improve the performance up to 5 percent more than others.

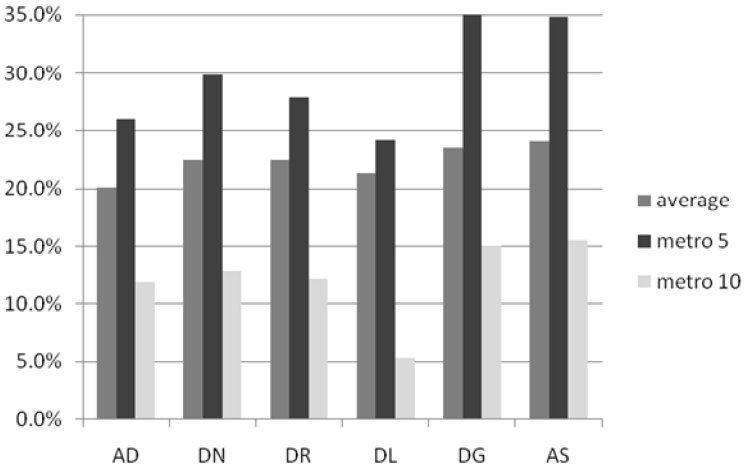


Fig. 7. Performance improvement by FSC.

IV. CONCLUSION

Quantifying similarity among regions in terms of fault characteristics is important for distribution fault diagnosis. In this paper, we proposed four different regional feature vectors: normalized regional feature vectors, relative regional feature vectors, likelihood regional feature vectors, and generalized regional feature vectors, all derived from measures used to analyze distribution faults. Similarity measures based on these regional feature vectors are defined and evaluated using fault events simulated by the Distribution Fault Simulator. Experimental results show us that:

- NRFV is the simplest measure among the four and represents the overall fault characteristics of the region disregarding the specific environmental factors. NRFV thus reflects the result of spatial, temporal and logical differences. Although DN is effective in representing the similarity, it could be confusing under complex situations. For example, a *rural* region with *nice* weather could be very similar to a *metro* region with *basic* weather in terms of DN, as found in Fig. 4.
- RRFV is good at representing the difference of spatial and temporal environmental information as it evaluates faults with the same root cause under different environmental conditions. Suppose the distribution lines are mostly overhead in one unit region and underground in another, then the relative measure of fault on overhead lines would be close to 1 in one region while close to 0 in another. However, changes in membership functions cannot be reflected by RRFV.
- LRFV reflects the logical difference as the fault events are simulated based on the likelihood of faults under given environmental conditions. Any change in the rules

and membership functions would directly affect the LRFV. LRFV can represent the difference in environments as well, but is only effective when the environmental difference is significant.

- GRFV combines the strength of the aforementioned three measures and performs the best in our evaluations. We would recommend using GRFV as the feature vector to represent the fault characteristics of a local region and using DG as the similarity measure.

Though the study of similarity measures is directly motivated by the need of identifying similar fault event sets in Small World Stratification, the findings reported in this paper are generally applicable to problems where similarity among different fault groups is of the interest. Thus, it would be helpful to other researchers and engineers in similar areas.

REFERENCES

[1] M.-Y. Chow and L. S. Taylor, "A novel approach for distribution fault analysis," *IEEE Trans. Power Delivery,* vol. 8, pp. 1882-1889, 1993.

[2] W.-H. Chen, C.-W. Liu, and M.-S. Tsai, "On-line fault diagnosis of distribution substations using hybrid cause-effect network and fuzzy rule-based method," *IEEE Trans. Power Delivery,* vol. 15, pp. 710-717, 2000.

[3] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Trans. Power Systems,* vol. 21, pp. 53-60, 2006.

[4] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm," *IEEE Trans. Power Systems,* vol. 22, pp. 164-171, 2007.

[5] L. Xu, M.-Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using Artificial Immune Recognition System (AIRS) algorithm," *IEEE Trans. Power Systems,* vol. 22, pp. 198-204, 2007.

[6] V. B. Nunez, S. Kulkarni, S. Santoso, and J. Melendez, "Feature analysis and classification methodology for overhead distribution fault events," in *IEEE Power and Energy Society General Meeting 2010,* Minneapolis, MN, 2010.

[7] Y. Cai and M.-Y. Chow, "A novel sampling strategy for distribution fault diagnosis: small world stratification," *IEEE Trans. on Power Systems,* submitted.

[8] Y. Cai and M.-Y. Chow, "Cause-effect modeling and spatial-temporal simulation of power distribution fault events," *IEEE Trans. Power Systems,* accepted.

[9] L. Xu, M.-Y. Chow, and L. S. Taylor, "Analysis of tree-caused faults in power distribution systems," in *the 35th North American Power Symposium*, University of Missouri-Rolla in Rolla, Missouri, 2003.

[10] Wind, [Online]. Available: http://en.wikipedia.org/wiki/wind.

[11] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics,* vol. 28, pp. 100-108, 1979.

[12] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2005, 2005.

[13] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Evaluation of distribution fault diagnosis algorithms using ROC curves," in *IEEE Power and Energy Society General Meeting 2010,* Minneapolis, MN, 2010.

CHAPTER VII

DESIGN OF SMALL WORLD STRATIFICATION ALGORITHM FOR DISTRIBUTION FAULT DIAGNOSIS

Yixin Cai ycai2@ncsu.edu

Mo-Yuen Chow chow@ncsu.edu

Department of Electrical and Computer Engineering

North Carolina State University

Raleigh, NC 27695 USA

This chapter is submitted to IEEE Transactions on Power Systems and is under review.

DESIGN OF SMALL WORLD STRATIFICATION ALGORITHM FOR DISTRIBUTION FAULT DIAGNOSIS

Abstract— Small World Stratification (SWS) is a sampling strategy that improves fault diagnosis performance when diagnosing distribution faults in a small local area. This paper details the algorithm design of SWS. We implement the process of sampling relevant fault events from other unit regions as detecting closely connected clusters in a fault event network. We first build a fault event network by adding geographic edges and similarity edges between fault event sets, and then identify closely connected clusters using community detection algorithms. Both simulated fault events and real-world outages from Progress Energy Carolinas' distribution systems are used to test the algorithm. Experimental results show that SWS effectively improves the fault diagnosis performance and, in general, is better than geographic aggregation.

Index Terms—community detection, discrete event simulation, fault diagnosis, power distribution faults, power system simulation, sampling methods, small-world.

I. INTRODUCTION

Small World Stratification (SWS) is a sampling strategy that improves the fault diagnosis performance when diagnosing distribution faults in a small local area. In our previous work, we proposed the concept of SWS and demonstrated its effectiveness with simulated fault

events [1] and discussed the use of the Euclidean distance between regional feature vectors as a metric for similarities between fault event sets [2]. However, a satisfactory method to identify proper clusters of fault event sets is yet to be discovered.

The idea of Small World Stratification is inspired by the small-world phenomenon first discovered in the study of social networks [3]. A small-world network is characterized by a small average shortest path length and a large clustering coefficient [4]. In other words, a small-world network is a network where nodes are relatively close to each other in the sense that only a small number of hops are needed to reach each other although the connections are globally sparse, and where local clusters formed by densely connected nodes exist. The small-world network has been used in the study of power systems in various perspectives. The early work mainly focused on characterizing the topology of power grids using small-world network models [4, 5]. The recent research by Wang, Scaglione and Thomas applied the small-world network model to constrain the topology of generated power lines for simulation purposes [6]. The vulnerability of power grids, i.e. whether the grid can sustain operation with failure of nodes, was investigated from the network connection perspective as well [7].

The SWS algorithm we propose in this paper is inspired by the research on community structures of small-world networks. We use community detection in a fault event network to implement the idea of SWS sampling, which identifies the relevant fault event sets that should be investigated together. The approach to build a fault event network and community detection in such a network will be discussed in detail.

The rest of this paper is organized as follows: Section II introduces the SWS algorithm, which includes building a fault event network by adding geographic edges and similarity edges between fault event sets and identifying closely connected clusters by community detection algorithms; Section III presents two case studies with both simulated and real-world fault events; Section IV is the conclusion.

II. Small World Stratification Algorithm

A. Algorithm Overview

The SWS algorithm implements the process of sampling relevant fault events as detecting clusters in a network formed by fault event sets. In the fault event network, each node represents a fault event set and the edges between nodes indicate that those fault event sets are relevant and may serve as the support set of each other. Corresponding to the two ways of sampling relevant fault events, Geographic Aggregation (GA) and Feature Space Clustering (FSC), two types of edges, geographic edges and similarity edges, are added between nodes when building the network. Closely connected clusters of nodes are then detected by community detection algorithms from the complex network literature. The overall algorithm steps are summarized in Fig. 1.

Detail on the algorithm design and implementation is explained in the following subsections with an example.

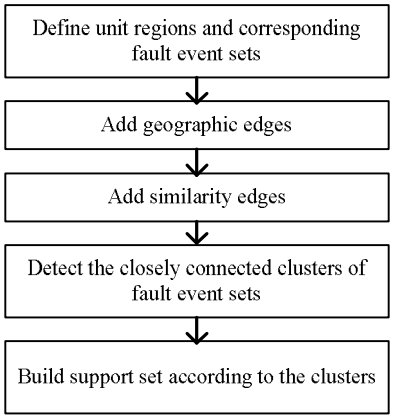


Fig. 1. Major steps of Small World Stratification algorithm.

B. Build a Fault Event Network

As defined in [1], a *unit region* is the smallest geographic area within which fault events are investigated as a group, and a fault event set is a set of fault events that occurred within a unit region:

$$s_j = \{e_i \mid (x_i, y_i) \in ur_j\}. \tag{1}$$

A fault event network is a network used to represent the relevance between fault event sets. Every node in the fault event network represents a fault event set and each edge indicates the two nodes it connects to are relevant in fault diagnosis.

Geographic edges are based on the geographic proximity of unit regions. Two unit regions ur_i and ur_j are geographic neighbors as long as they share one piece of common boundary, denoted by $ur_i \sim ur_j$. Correspondingly, a geographic edge is added between fault event sets s_i and s_j if the unit region ur_i and ur_j are geographic neighbors, represented by:

$$eg(s_i, s_j) = \begin{cases} 1 & if\ ur_i \sim ur_j \\ 0 & otherwise \end{cases}. \tag{2}$$

Suppose there are two service areas consisting of several *metro* type unit regions (model of typical distribution systems in cities and major town, where the fault event sets are denoted by *M*) and *rural* type unit regions (model of typical distribution systems in rural areas, where the fault event sets are denoted by *R*), as shown in Fig. 2. The fault event network after adding geographic edges is shown in Fig. 3. Note that this network is a direct mapping of the map in Fig. 2, where fault event sets fall in two separate subnets. Every fault event set has between 2 to 5 neighbors and the neighbors could be of the same or different type.

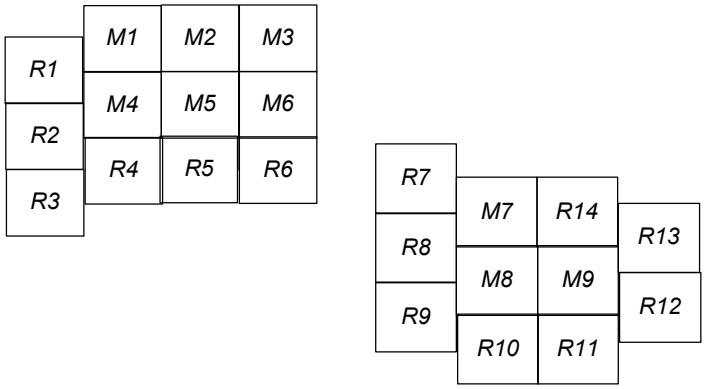


Fig. 2. An example of two service areas consisting of multiple unit regions.

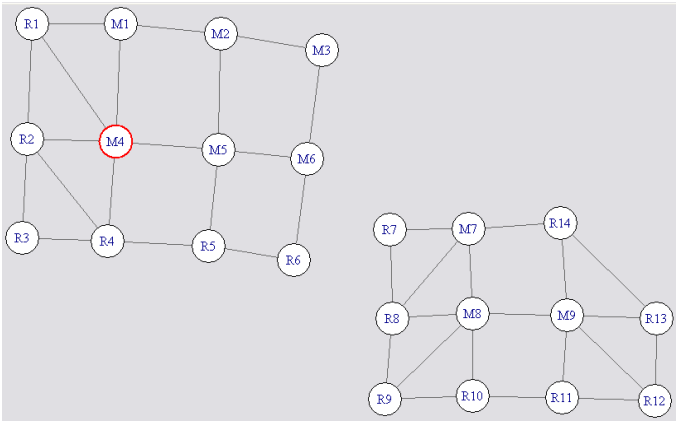


Fig. 3. The fault event network by adding geographic edges.

Similarity edges consider the fault characteristics of a fault event set. Suppose the fault characteristics of the fault event set s_i can be represented by a feature vector $\mathbf{F}_i$, and the difference between fault characteristics can be measure by a distance function $D(\mathbf{F}_i, \mathbf{F}_j)$. Based on our study on similarity measures [2], we choose the Generalized Regional Feature Vectors (GRFV) to represent the fault characteristics and the Euclidean distance between GRFVs as the distance measure in this paper. Similarity edges are added between two fault event sets s_i and s_j when the distance in the feature space is less than a predefined threshold d_0:

$$ef(s_i, s_j) = \begin{cases} 1 & D(\mathbf{F}_i, \mathbf{F}_j) < d_0 \\ 0 & D(\mathbf{F}_i, \mathbf{F}_j) \geq d_0 \end{cases}. \quad (3)$$

For the service areas shown in Fig. 2, we get a network as shown in Fig. 4 by adding similarity edges to its nearest 3 neighbors of every node in the feature space. We can see that this network has two separate subnets representing the two different types of fault event sets and the most connected fault event set has 8 neighbors.

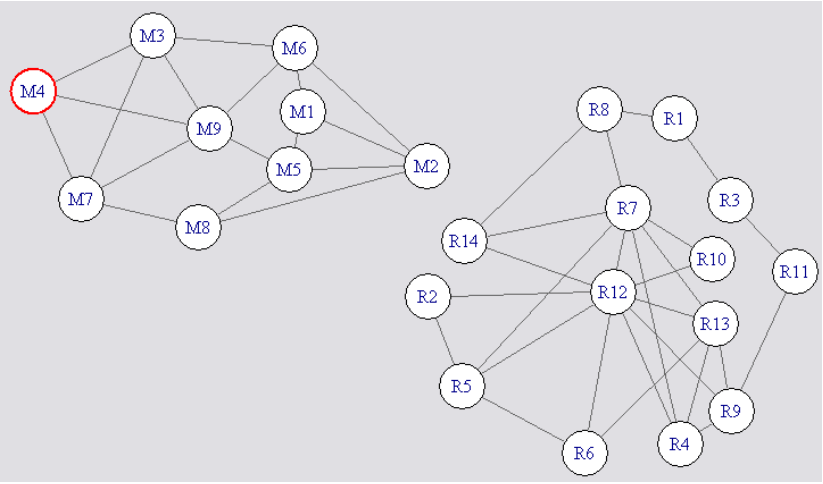


Fig. 4. The fault event network by adding similarity edges.

The two-layered view shown in Fig. 5 better illustrates the way of building a fault event network in the SWS algorithm. In the geographic space, fault event sets are connected as two subnets based on the locations of the corresponding unit regions. When mapping into the feature space (represented by the dashed lines. Only the mapping of the *M* type fault event sets is drawn.), fault event sets from the same type of unit regions are connected and the network is divided into two subnets as well. Each layer provides information regarding how relevant the fault event sets are from a certain perspective and the information could be different. With a real-world dataset, the fault event network usually is not clearly separable as in this example. In either case, we will combine these two layers to form a connected network and analyze the edges among nodes to find the closely connected clusters.

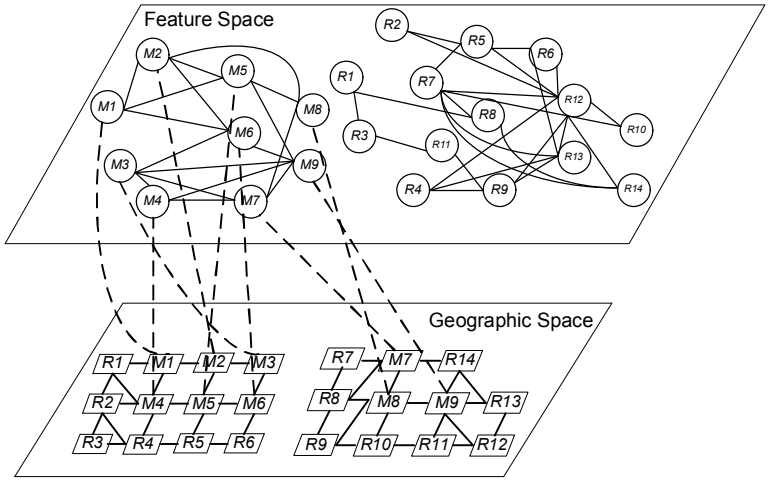


Fig. 5. The two-layered view of the fault event network.

C. Find Clusters by Community Detection

Community detection is a topic of great interest in the complex network literature. Communities, also called clusters, are groups of nodes which probably share common properties and/or play similar roles within the network [8, 9]. Usually, communities are

characterized by many edges among nodes within the group and few edges between groups. Algorithms optimizing different measures, such as modularity [10] and conductance [11], to detect communities in networks have been proposed and compared. For a comprehensive review of community detection, please refer to [8].

Detecting clusters in a network can be viewed as dividing the network in such a way that each subnet is a closely connected cluster. Modularity is a measure proposed by Newman and Girvan [12] to evaluate the relevance of a particular division. The fraction of all edges that lie within communities is first calculated. The expected value of the same quantity in a network where the nodes have the same degrees but edges are placed at random without regard to communities is used as a reference. Modularity is defined to be the difference of these two numbers. The larger modularity, the better the division is. A modularity of zero means the division of communities is no better than a random guess. An algorithm greedily searching the edge removal that leads to maximum modularity increase is known as CNM [10]. The final division of the network is chosen as the one that maximizes the modularity value.

Community detection algorithms provide a potential tool to find closely connected clusters of fault event sets. However, whether the clusters reflect reality depends on the way we build the network. Through experiments, we observed that:

- Quite a number of duplicated edges exist. Geographic edges and similarity edges are generated independently. There are duplicated edges between two nodes when they are geographic neighbors and feature space neighbors at the same time. In the feature space, if the nearest neighbor of s_i is s_j, s_j is not necessarily the nearest neighbor of s_i.

However, when it is, the duplicated similarity edges between s_i and s_j indicate stronger relevance than a single edge. Thus, although we need to remove duplicated edges to keep the network a simple undirected one, the number of duplicated edges are counted and used as the weight of the simplified edge.

- The number of similarity edges matters. Unlike the definition of geographic edges, there is a parameter d_0 in the definition of similarity edges, which means the number of similarity edges is up to the design. If the number of similarity edges is too big, the entire network is densely connected so that it appears as a whole without community structures. The number cannot be too small either. Otherwise the geographic edges will dominate the community structure while the similarity in fault characteristics will not affect much. Through experiments with simulated fault event sets, we find an empirical rule for the number of similarity edges. The threshold d_0 is selected such that

$$\frac{\#\,similarity\ edges}{\#\,geographic\ edges} = \frac{\#\,geographic\ edges}{\#\,nodes}. \tag{4}$$

For example, the network shown in Fig. 5 has 23 nodes and 38 geographic edges. According to (4), 63 similarity edges are added. The simplified network has 73 edges in total with 12 weighing 3 and 38 weighing 2, as shown in Fig. 6.

We apply the CNM community detection algorithm to the fault event network shown in Fig. 6. Although the edges are too busy for human eyes to detect a pattern, the algorithm finds three communities and separates the *M* type fault event sets from the *R* type. With small

groups within the *R* type fault event sets, the clusters detected generally reflect the actual situation.

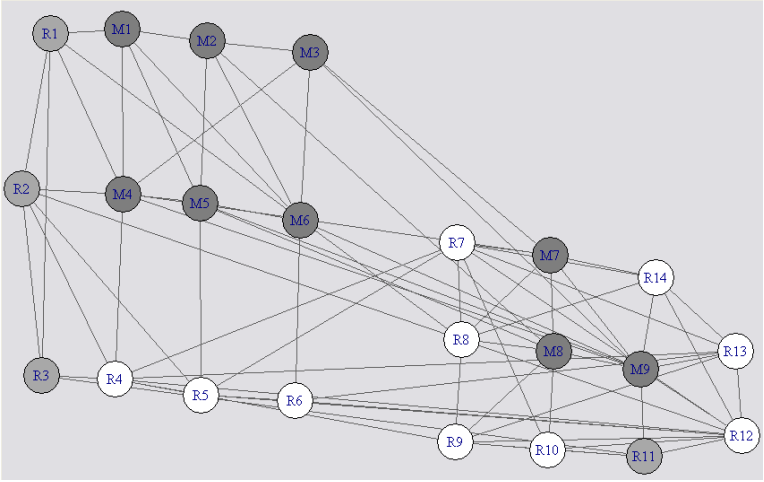


Fig. 6. The fault event network and the three clusters detected.

III. CASE STUDIES

A. Case 1: Simulated Fault Events

In this case study, we test the proposed SWS algorithm on a simulated fault dataset generated by the Distribution Fault Simulator [13]. Suppose we have collected historical fault events from 6 service areas consisting of different numbers of *metro* and/or *rural* unit regions, as shown in Fig. 7. Two of the service areas are relatively big, with *metro* unit regions surrounded by *rural* unit regions, which represent the typical distribution systems serving an urban area and its surroundings. Other small service areas contain 2 to 4 unit regions of the same type, representing local distribution systems with limited data collection capabilities. The challenge for the SWS algorithm is whether it can distinguish the two different types of

fault event sets in the big service areas and whether it is able to find relevant fault event sets beyond geographic neighbors for the small service areas.

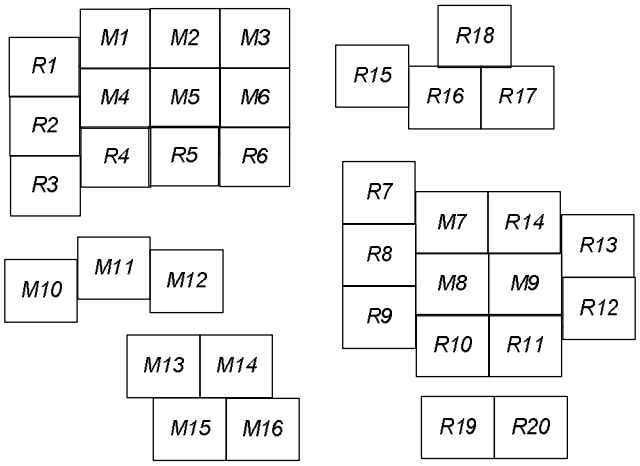


Fig. 7. Case Study 1: simulated fault events from multiple service areas.

The fault event network built in this case has 36 nodes, 50 geographic edges and 70 similarity edges. Four clusters are detected by CNM as represented by different shades in Fig. 8. We can see that *M* type fault event sets are well separated from *R* type and the support set for every fault event set could go beyond its geographic neighbors. This is especially important for small service areas such as the one consisting of *R19* and *R20*, where fault events from the geographic neighbor are very likely to be insufficient.

We evaluate the improvement of fault diagnosis performance as follows:

- Take every fault event set in this case study as the target set. The support set for each target set is the union of all fault event sets belonging to the same cluster as the target set.
- Logistic Regression model is trained on the support set in addition to half of the target set in order to diagnose tree-caused faults from non-tree faults.

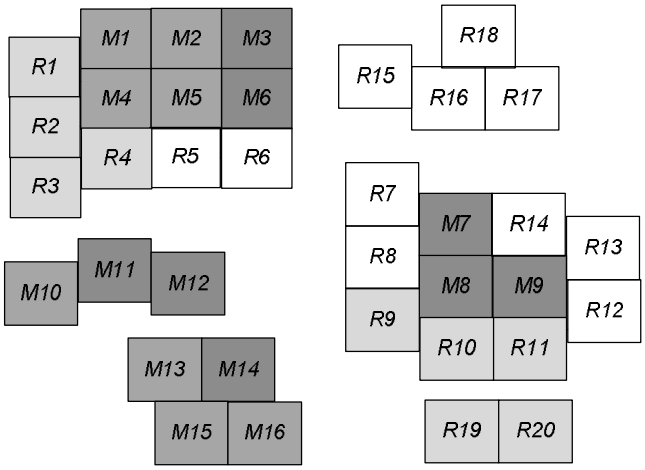


Fig. 8. Clusters detected for Case Study 1.

- The fault diagnosis performance is evaluated by the G-mean measure [14] and the decision threshold is selected to maximize the training G-means. The LR model is then tested on the remaining half of the target set.

The average testing G-means for diagnosing faults based only on the target set (TO), sampling by geographic aggregation (GA) and sampling by SWS (SWS) are compared by two-sample *t*-test at 0.05 significance level. The result is summarized in Fig. 9.

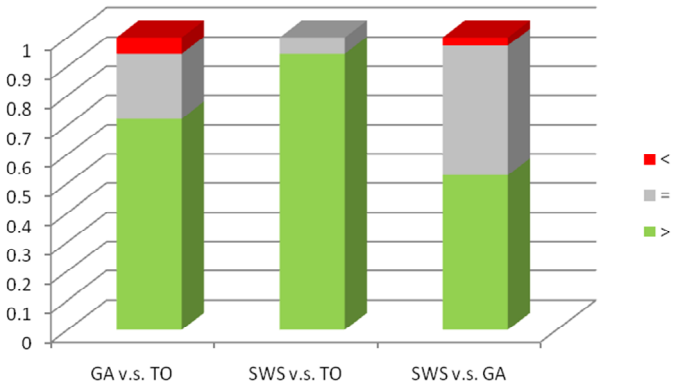


Fig. 9. Comparison of target set only (TO), geographic aggregation (GA) and small world stratification (SWS) by *t*-test.

The middle bar in Fig. 9 shows that SWS has superior performance (green) to TO on more than 90% of target sets and equivalent performance (gray) on less than 10% of the target sets. Generally, GA has superior performance to TO but can have inferior performance (red) on occasion. SWS is comparable to or superior to GA more than 90% of the time. So in general we may conclude that SWS is effective in improving the fault diagnosis performance and is a better choice than GA.

B. Case2: Real-World Fault Events

The dataset used in this case study is from the outage database of Progress Energy Carolinas. Outages between 2005 and 2006 in the Garner operation center are studied (as shown in Fig. 10).

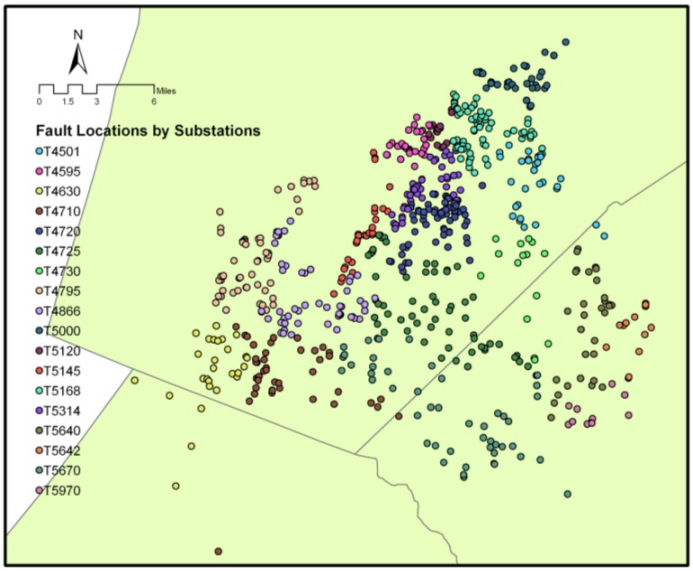


Fig. 10. Fault events from Garner operation center by substations.

Unit regions in this case are defined as the service territory of a distribution substation. Eighteen unit regions are identified from the fault event records. The number of available fault events in each unit region ranges from 12 to 78, as shown in Fig. 11. Based on our

experience with the simulated fault events, none of these unit regions contains sufficient historical data for the fault diagnosis algorithm to make proper inference.

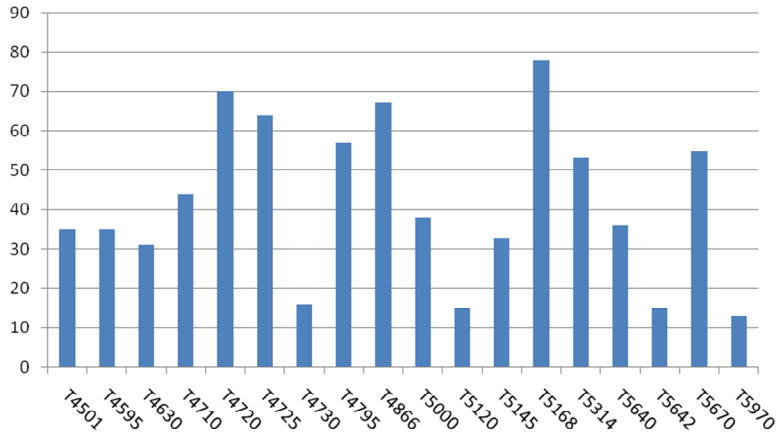


Fig. 11. Number of historical fault events in each substation.

As the boundary of substation service territory is not well-defined in the outage database, the geographic neighbors of each unit region are identified by observing the fault event location map in Fig. 10. Thirty-six geographic edges are added, as represented by the solid line in Fig. 12. The same GRFVs consisting of 66 components as in [1] is used to represent the fault characteristics of each unit region. According to (4), 72 similarity edges are added to the fault event network. To keep the figure legible, only a small part of the similarity edges are drawn as dashed lines in Fig. 12.

Two clusters are detected by CNM as represented by shades in Fig. 12. Comparing to the map, the shaded unit regions roughly corresponds to the rural area while the white unit regions covers mostly the metro part of Garner and the southern part of the city of Raleigh.

Similar to Case 1, the average testing G-mean of diagnosing faults based only on the target set, sampling by geographic aggregation and sampling by SWS are summarized in Fig. 13 and the result of two-sample *t*-test at 0.05 significance level is shown in Fig. 14.

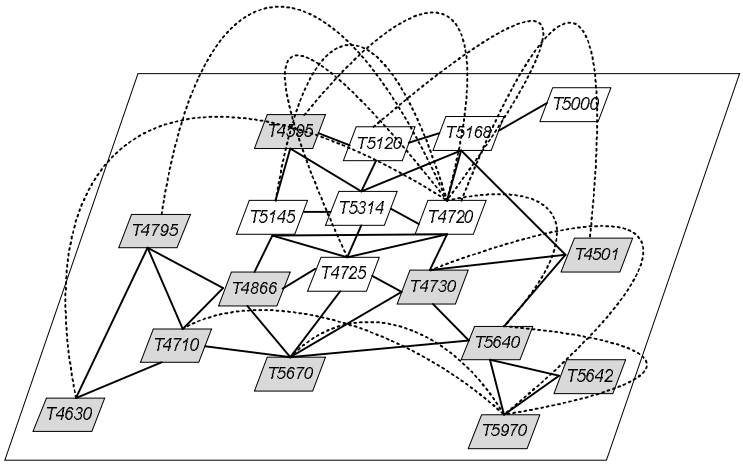


Fig. 12. The fault event network and clusters detected for Case 2.

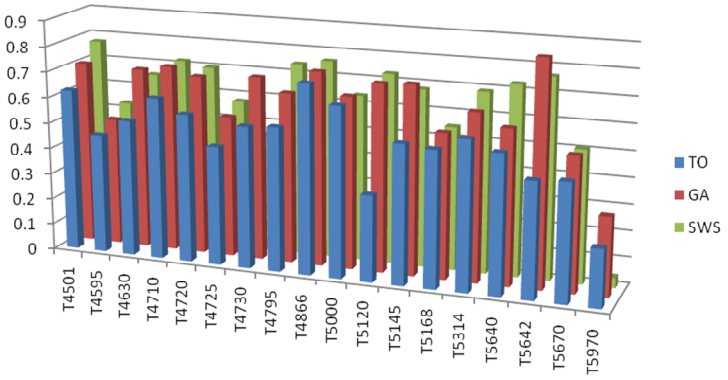


Fig. 13. Average testing G-means for Case 2.

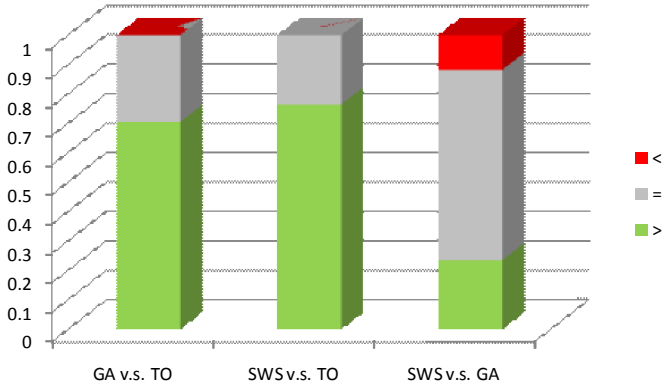


Fig. 14. Comparison of target set only (TO), geographic aggregation (GA) and small world stratification (SWS) by *t*-test.

From Fig. 13 and 14 we can see that both GA and SWS are capable improving fault diagnosis performance. However, with the real-world fault events, the difference between GA and SWS is not that obvious. One possible reason is that the simulated *M* and *R* type fault event sets are quite different as they are supposed to model two typical systems. As we discovered in the previous demonstration [1], SWS is superior to GA when the neighboring unit regions are of different fault characteristics. While in this case study, the difference among neighboring unit regions is not as significant as the simulation case so GA and SWS are mostly comparable.

IV. CONCLUSION

This paper focuses on the algorithm design and implementation of Small World Stratification. Inspired by the research on community structures of small-world networks, we implement the process of sampling relevant fault events from other unit regions as detecting closely connected clusters in a network formed by fault event sets. A fault event network is first built by adding geographic edges and similarity edges between fault event sets and the clusters are then identified by the CNM community detection algorithm. Both simulated fault events and real-world faults from a local distribution system of Progress Energy Carolinas are used to test the SWS algorithm. Experimental results show that SWS effectively improves the fault diagnosis performance and is generally better than geographic aggregation.

V. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of John W. Gajda and Glenn C. Lampley from Progress Energy Carolinas Inc. for their support on data and field experience.

REFERENCES

[1] Y. Cai and M.-Y. Chow, "A novel sampling strategy for distribution fault diagnosis: small world stratification," *IEEE Trans. Power Systems,* submitted.

[2] Y. Cai and M.-Y. Chow, "Measuring similarity among regions for distribution fault diagnosis," *IEEE Trans. Power Systems,* submitted.

[3] S. Milgram, "The small world problem," *Psychology Today,* vol. 2, pp. 60-67, 1967.

[4] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature,* vol. 393, pp. 440-442, 1998.

[5] P. Crucitti, V. Latora, and M. Marchiori, "A topological analysis of the Italian electric power grid," *Physica A,* vol. 388, pp. 92-97, 2004.

[6] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating statistically correct random topologies for testing smart grid communication and control networks," *IEEE Trans. Smart Grid,* to be published.

[7] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the North American power grid," *Physical Review E,* vol. 69, pp. 1-4, 2004.

[8] S. Fortunato, "Community detection in graphs," [online]. Available at: http://arxiv.org/abs/0906.0612.

[9] M. E. J. Newman, "Detecting community structure in networks," *the European Physical Journal B - Condensed Matter and Complex Systems,* vol. 38, pp. 321-330, 2004.

[10] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E,* vol. 70, 2004.

[11] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Statistical properties of community structure in large social and information networks," in *WWW 2008* Beijing, China, 2008.

[12] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E,* vol. 69, 2004.

[13] Y. Cai and M.-Y. Chow, "Cause-effect modeling and spatial-temporal simulation of power distribution fault events," *IEEE Trans. Power System,* accepted.

[14] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Evaluation of distribution fault diagnosis algorithms using ROC curves," in *IEEE Power and Energy Society General Meeting 2010,* Minneapolis, MN, 2010.