Paper accepted for presentation at 2009 IEEE Bucharest Power Tech Conference, June 28th - July 2nd, Bucharest, Romania

Data-Driven Reliability Modeling, Based on Data Mining in Distribution Network Fault Statistics

E. Akhavan-Rezai, *Student Member*, *IEEE*, M. -R. Haghifam, *Senior Member, IEEE*,
and A. Fereidunian, *Member,* IEEE

***Abstract*—Power distribution fault statistics provide splendid resource for extracting experimental knowledge. The extracted knowledge includes the inherit characteristics of the network assets. Analysis and estimation of failures require a comprehensive understanding of faults in terms of the relevant effective parameters.**

This paper outlines a data-driven model to represent momentary failure rate in terms of the most influential factors based on the study of the recorded historical fault data as well as the expert's experiments in the Greater Tehran Electricity Distribution Company. A methodology is presented for momentary fault causes identification and model construction using artificial neural networks. Satisfactory results indicate that the developed model can easily be implemented to estimate other fault types in power distribution systems.

***Index Terms*—Artificial neural networks, distribution system reliability, data mining, fault cause identification, failure-rate estimation, failure-rate modeling, historical data analysis.**

I. Introduction

Maintaining and improving the reliability of power distribution systems has become a core challenging activity for electric utilities, due to deregulation and privatization of the utility industry. As a result, utilities –who strive to improve their reliability indices–, need to improve their reliability estimation and monitoring methods. For this reason, electric utilities have developed modern tools for optimal asset management, based on distribution network reliability analysis. Fault statistics –treasured in IT-based management automation systems of the electric utility companies– provide splendid resources for extracting experimental knowledge. This knowledge extraction is referred to as *Data Mining* or *Knowledge Discovery* [1]. The extracted knowledge include the inherit characteristics of the network asset engineering practice within that company, which might be exclusive to its specific network, thus very valuable for that utility. Hence, the utilities record and analyze their fault statistics; as they would rather prefer to rely on their own context-specific extracted knowledge, rather than the typical values [2, 3]. Therefore, the utilities need practical and adaptive reliability indices, based on their own data, in order to direct their resources properly [3].

Analytical models often use simplifying assumptions, such as constant failure rates for systems components, in order to cope with complexities [4]. Since the use of average failure rates in system reliability models is always limiting and is potentially misleading, advanced tools have attempted to move beyond average failure rates by calibrating or modeling it based on historical system performance [5]. A few attempts have been made to evaluate failures as a function of either external factors, such as plant growth [6, 7], weather [8], combination of features [4, 9], or intrinsic factors, such as age [10]. A method that uses equipment inspection data to assign relative condition ranking was addressed in [5], these rankings then mapped to a failure rate function. A model by dividing the component failure rates into several partial failure rates are introduced in [11]. Attempts have also been devoted to allocate failures to improve repair activities [12].

In spite of the great efforts made in former studies, this research area is still young. In fact, vast varieties of failure causes, as well as the effects of indirect environmental/structural factors on equipments failures are considerably significant. Hence, more efforts on this issue could provide valuable reliability knowledge and practical guidelines. As a result, the necessity of more capable models to adapt reliability assessments to different environmental conditions is inevitable.

This paper is aimed to represent a data-driven artificial neural networks model for momentary failures of overhead medium voltage lines in terms of most influential factors, based on the study of the recorded historical fault data as well as the expert's experiments in the Greater Tehran Electricity Distribution Company (GTEDC). Section II includes a brief introduction of the GTEDC service territory. The proposed methodology and the data mining approach including analysis of the historical outage data are presented in sections III and IV respectively. Section V explains the failure rate model; and the implementation results, model evaluation, as well as the discussion on the results are shown in section VI.

This research has been performed in collaboration with the Greater Tehran electricity Distribution Company (GTEDC), Tehran, Iran.

Elham Akhavan-Rezai and Mahmoud-Reza Haghifam (corresponding author) are with the Department of Electrical Engineering, Islamic Azad University, Tehran-South Branch, Tehran, Iran.

Alireza Fereidunian is with the CIPCE, School of ECE, University of Tehran, P.O.Box 14395-515; and serving as an independent consultant to the GTEDC, Tehran, Iran.

II. The Greater Tehran Electricity Distribution Network

GTEDC operates the distribution network of The Greater Tehran, the capital city of Iran. It is the largest electric utility in the country, providing almost 16641[GWh/ year], 3249000 customers, and 13300 MV substations [13]. Like many other utilities, GTEDC is continually making an effort to improve modeling and estimation methods for reliability indices. It is divided into four regional areas. Since the environmental conditions vary within different zones of the city, it could significantly affect the system reliability indices. A geographical network map of the four GTEDC's service areas as *North West (NW), North East (NE), South West (SW), and South East (SE)* is illustrated in Fig. 1.

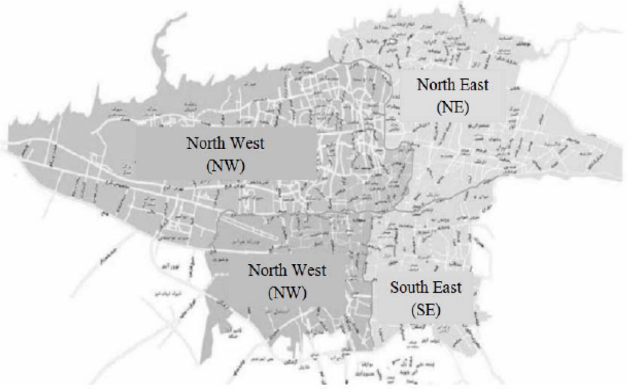

Fig. 1. Greater Tehran geographical network map including four service areas as *North West (NW), North East (NE), South West (SW), and South East (SE).*

III. Methodology

In this study, a data-driven model is used to estimate momentary failure rates based on fault statistic in the outage management database. As shown in Fig. 2, the proposed methodology consists of two main stages: data mining and failure rate modeling.

In the data mining stage, the historical data of the network momentary interruptions are analyzed first. Then, the fault data are preprocessed, and the momentary failure cases are identified, using a statistical data-mining approach.

The failure rate modeling stage, subsequently, starts with preparation of a data-driven pattern recognizer model; followed by introducing the finally designed alternative models. The alternative models should be evaluated to ensure the satisfactory performance of the designed model, as well as to select one of the alternative models as the proposed reliability estimation model.

IV. Data Mining

A. Analysis of the Historical Data

In order to gain an insight on failure events and assess different regional factors, four districts of the GTEDC were studied in details. The fault statistics data are registered in *Electric Network Operating eXpert* (ENOX^{TM}), a home-developed relational outage management database. Like other outage management systems [4,7,9], it records various fields of outage information, including almost fifteen items, such as short event description, fault location and circuit ID, failure cause, repair duration and weather situation.

B. Fault Data Preprocessing

First step in failure rate model extraction is study of unscheduled historical outage events. Analysis was done using recorded historical outage data of overhead medium voltage lines in ENOX^{TM} data base for the period of 2006-2008. This includes more than 4900 unscheduled outage events on overhead medium voltage lines. Interruption frequency of the different failure types and the contribution of each region on failures are shown in Figs. 3 and 4, respectively. As shown in Fig. 3, the dominant failure type of overhead lines is momentary type including almost 58% of all of the total failure numbers (2808 out of 4906). Fig. 4 shows that the momentary interruptions are distributed within the four service areas of GTEDC. Since the network includes insufficient local automation equipments (reclosers), momentary interruptions last around 15 or 20 minutes. Consequently, it yields to about 35% (696 MWh out of 2038 MWh) of the total annual energy not supplied (ENS) of overhead lines, as illustrated in Fig. 5.

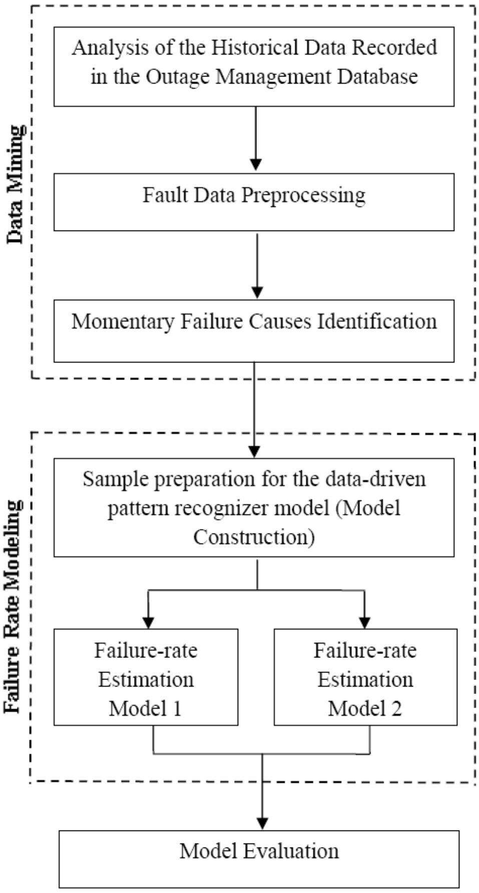
Data Mining
Analysis of the Historical Data Recorded in the Outage Management Database
Fault Data Preprocessing
Momentary Failure Causes Identification
Failure Rate Modeling
Sample preparation for the data-driven pattern recognizer model (Model Construction)
Failure-rate Estimation Model 1
Failure-rate Estimation Model 2
Model Evaluation

Fig 2. Flowchart of the proposed methodology for failure rate modeling.

On the other hand, momentary interruptions are more unknown in nature than sustained interruptions; as, the feeder can be re-energized right after a momentary event without any repair action. Therefore at the first phase of our research we focused on momentary faults and we postpone other types.

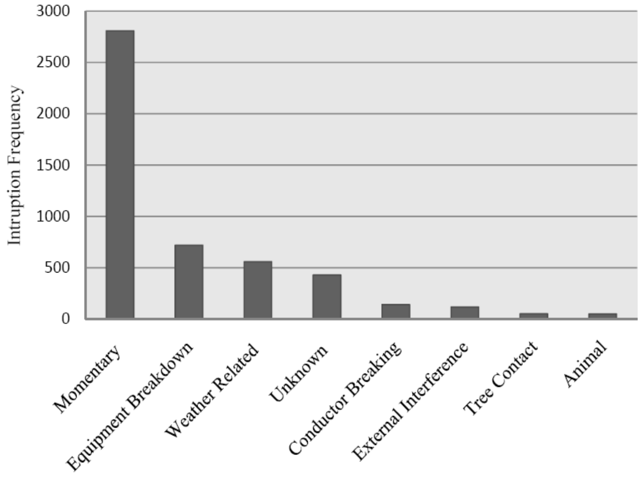

Fig. 3. Interruption frequency for different types of outages in overhead lines of the GTEDC service territory for the period of 2006-2008.

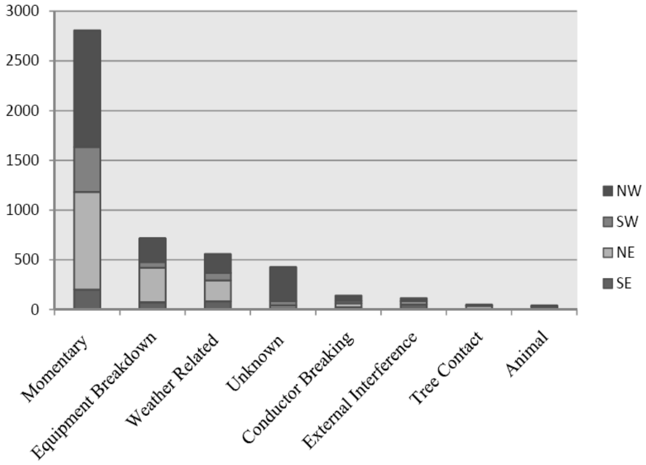

Fig. 4. Contribution of each area of the GTEDC on interruption frequency for different outage types.

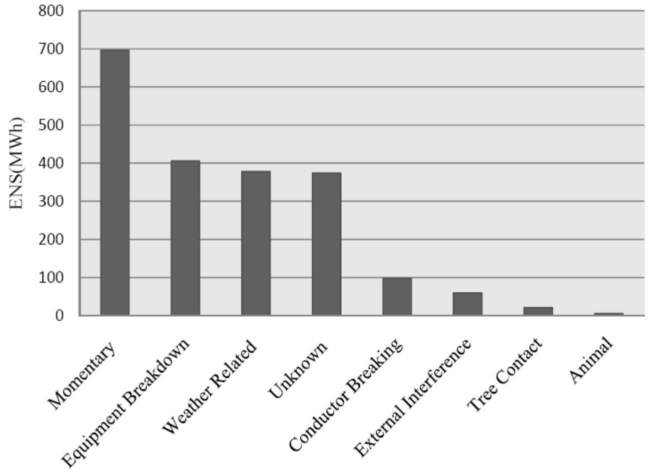

Fig. 5. Annual average ENS (MWh) versus different failure types.

C. *Momentary Failure Causes Identification*

Failures are functions of various independent parameters, thus, the identification of causes -and their associated consequences- is one of the most formidable challenges of reliability studies as well as preventive maintenance strategies. According to [3], various factors that cause failures in distribution systems can be divided into three main groups, as *External factors*, such as trees, birds/animals, wind, etc, *Intrinsic factors*, such as age of equipment, manufacturing defects, size of conductors and *Human error factors*, such as vehicular accidents, accidents by utility or contractor work crew, vandalism, etc.

Firstly, we explored the impacts of weather parameters as the influential factors on momentary fault types. By analysis of adverse weather days, it was found that there are no significant correlation between weather conditions and momentary failures. This includes correlation analysis of historical data of daily wind speed and rainy/snowy days collected from the weather meteorology substations for the period of 2006-2008, along with the relevant historical outage data. Although, effects of weather conditions on fault occurrence are inevitable in general; however, this result is valid and expectable, since in this area weather is usually mild and there were few adverse weather days that influence on momentary faults in a period of the study. The primary result led us to go beyond the weather parameters and focus on identification of more effective probable factors.

Consequently, we faced with a fundamental technical challenge: the ENOX^{TM} database does not report the main momentary fault causes. We talked this challenge as utilizing the experiences and observations of the GTEDC experts. After extensive studies, based on experts' guidelines on 160 overhead medium voltage feeders of the GTEDC, we found that outages are derived from four different classes of independent variables, as: *Poor Network Design and Installation, Inappropriate Operation Condition, Weather Condition,* and *External Environmental Interference.* Table I summarizes this classification including relevant subsets and the correlation coefficient of each variable with the total annual momentary failures.

V. Failure Rate Modeling

Due to complexities of the problem space we used artificial neural networks (ANN), as a model-free approach to cope with nonlinear failure rate modeling. ANN have been used successfully in the past for different purposes such as, fault detection, control, pattern recognition [14, 15], and fault cause identification [16]. One of the most common/famous network is a feed-forward network witch is trained with back propagation algorithm [17]. This algorithm has the ability to generate beyond the training data [6]. After extensive experimentations, an ANN model with one hidden layer, and a binary sigmoid activation function yield the best results. The architecture of the applied ANN failure rate model is shown in Fig. 6.

TABLE I. MOMENTARY FAILURE CAUSES CLASSIFICATION

	Main Momentary Fault Causes	Correlation Coefficient
Poor Network Desighn and Installation	Length of the Feeder	0.6543
	Inappropriate Span Length	0.6157
	Inappropriate Line Sag	0.652
Improper Operation Condition	Extra MV Substations per feeder	0.5297
	Insufficient Tree Triming	0.4861
	Overloading	0.0641
	Improper Scheduled Preventive Maintenance	0.4482
	Exposure to Mechanical/Electerical Stress	0.0596
	Normal Aging	0.4726
Adverse Weather	Wind Speed	0.0853
	Storm	-
	Rainfall	0.0289
External Interactions	Animal Contact	0.261
	Human Interference	0.0539

In order to determine the problem parameters and prepare useful data samples for the model, level condition of all the 160 feeders were evaluated based on each mentioned factors in Table I. The applied ANN model was constructed with 8 input variables including valuable and informative data for each sample. These variables were selected according to the relevant correlation of more than 0.1 with momentary failures. Each input variable and its symbol are shown in Table II. This data is then normalized to be valid for the model.

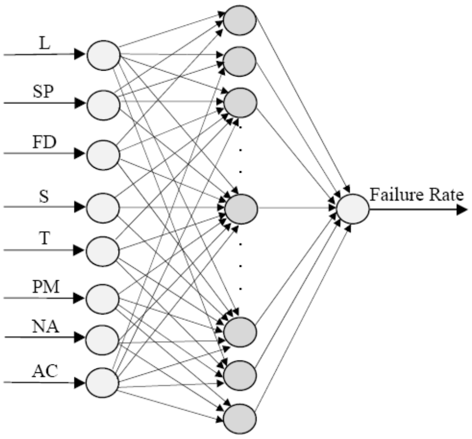

Fig. 6. Architecture of the applied ANN model.

TABLE II. INPUT VARIABLES FOR THE ANN MODEL

Symbol	Input Variables
L	Length of the Feeder
Sp	Inappropriate Span(Section) Length
FD	Inappropriate Line Sag
S	Extra MV Substations per Feeder
T	Insufficient Tree Triming
PM	Improper Scheduled Preventive Maintenance
NA	Normal Aging
AC	Animal Contact

VI. RESULTS AND DISCUSSIONS

A. Implementation Results

Here the results of the extracted model are presented. The finally designed failure rate estimation model will be an ANN with one hidden layer and a binary sigmoid activation function. Furthermore, the following data are used to train the ANNs: outage data from 2006 to 2008 and the condition of overhead medium voltage feeders according to the proposed failure case classification (Table I and Table II).

Seventy percent of failure data for each district was randomly selected for training, whereas the remaining thirty percent of the data was used to verify the generalization and the robustness of the tuned ANN model.

As mentioned before, momentary failures are more unknown in nature than sustained failure types. In order to assess the model, we evaluated the model performance sensitivity to changing the input variables. Therefore, two different models were developed separately. The first model includes 8 input variables as mentioned in Table II, whereas in second model 5 variables were selected out of 8, which have highest correlation coefficients. After reaching to the best estimation the root mean-square error (RMSE) for the training and test data sets of both models were determined using Equation (1). Table III summarizes the results of the final models.

$$\sqrt{\frac{\sum(\lambda_E-\lambda_o)^2}{n}} \qquad (1)$$

Where λ_E, λ_O, and n stand for expected failure, observed failure, and number of samples respectively.

TABLE III. RESULTS OF THE FINAL MODELS

Number of Inputs	Exact failures of test set	Observed failures of test set	Training set RMSE	Test set RMSE
8	468	484	0.0778	0.0821
5	468	491	0.005	0.1081

B. Evaluation

This study also evaluates the proposed method by comparing estimated failure numbers of test data set with the expected values, which were recorded in the ENOX^{TM} database. Fig. 7 depicts the exact (real) failure numbers versus the observed failure numbers, for 30 test samples. According to the results, first model including 8 input variables outperforms the second one.

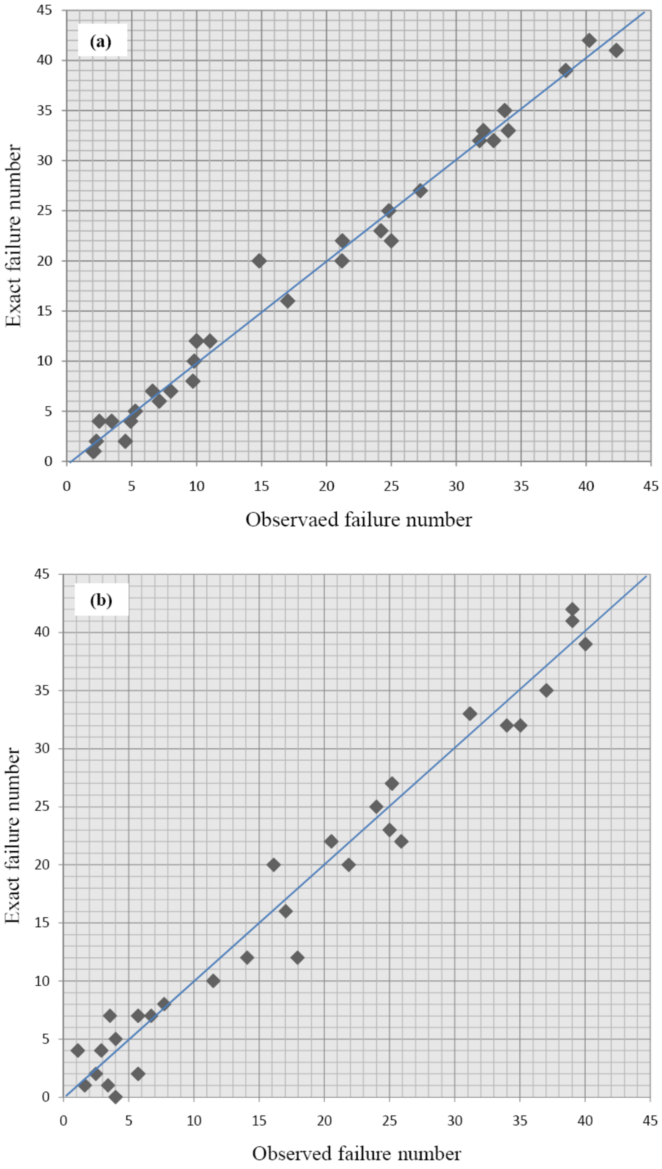

Fig. 7. The observed failure numbers versus exact failure numbers for 30 samples of test set, (a)8- input variable model, (b)5- input variable model.

C. Discussions

Evaluation of the presented models shows their satisfactory learning ability. It successfully overcomes the difficulties due to data shortage and uncertainty, which is a common problem in historical data analysis of utility databases.

The test to train error ratio of first and second models are 1.14 and 1.84 respectively (see Table IV). According to this ratio, it can be concluded that the first model (8-input ANN) has the better generalization capability than the second one (5-input ANN).

TABLE IV. RESULTS OF THE FINAL MODELS

Models	Generalization Capability	Training set RMSE	Test set RMSE
8-input ANN	1.143	0.0778	0.0821
5-input ANN	1.83	0.005	0.1081

The proposed model enjoys an evolutionary feature, since the data-driven characteristic of the model improve its performance as data samples gradually would increase in future.

The presented model is also useful for the utility managers to direct existing resources properly. As Reliability Centered Maintenance (RCM) is a strategy where maintenance of system components is related to the system reliability improvement; consequently, realistic evaluation of the network reliability conditions can be regarded as a start point of RCM programs. This RCM programs may include identification of weak parts/features of the system, as well as unknown failure factors.

VII. CONCLUSION

This paper advanced the premise of using context-specific and adaptive reliability indices, instead of typical values. The main momentary failure causes were identified, as the dominant type of failures, based on the study of the network using experiments of the GTEDC's experts.

This analysis followed by a classification of the more effective causes as *Poor Network Design and Installation, Inappropriate Operation Condition, Weather Condition, and External Environmental Interference*. Subsequently, a data-driven model was constructed to represent momentary failure rate in terms of the most influential factors. Satisfactory results demonstrate the proper performance of the proposed model.

In future, as the next phase of this study we are going to use these results to focus on improving the weak parts of the network. This can be regarded as helpful decision aids for planning future asset management and preventive maintenance policies. The results can also help us on improving the outage management database (ENOX^{TM}) with Failure Estimation System as well as data warehousing enhancement. Furthermore, adaptive and accurate estimates for failure rates, lead us to better evaluation of system reliability indices such as MAIFI and SAIFI.

VIII. ACKNOWLEDGMENT

This research has been performed in collaboration with the Greater Tehran Electricity Distribution Company (GTEDC); in which, the instrumental cooperation of the following colleagues are thankfully acknowledged: Mr. Payam Kharazmi (Senior Engineer), Mr. H. Torabi (CEO Deputy), Mr. A. Kazembakhshi (Technical Deputy), and GTEDC's

SCADA experts, especially, Mr. H. Mozaffari, Mr. Kh. Torangan, and Mr. M.H. Bahrami.

The authors would also like to thank to Mr. Mohammad-Ali Zamani (of CIPCE, School of ECE, University of Tehran) for his cooperation in modeling and training of the ANNs.

IX. References

[1]- Larose, D.T., 2004,"Discovering Knowledge in Data: An Introduction to Data Mining", Wiley Co.

[2]- PYLVÄNÄINEN, J., VERHO, P., JÄRVINEN, J., 2005, "Advanced Failure Rate and Distribution Network Reliability Modeling as Part of Network Planning Software", Proceedings of the CIRED 2005, 6-9 June, Turin, Italy.

[3]- Pahwa, A., 2004, "Effect of Environmental Factors on Failure Rate of Overhead Distribution Feeders" in *Proc*. IEEE PES General meeting, pp. 691-692.

[4]- Gupta, S., Pahwa, A., Brown, R.E., Zhou, Y., Das, S., 2005, "An Adaptive Fuzzy Model for Failure Rates of Overhead Distribution Feeders", Electric Power Components and Systems, pp. 1175– 1190.

[5]- Brown, R.E., Frimpong, G., Willis, H.L., 2004, "Failure Rate Modeling Using Equipment Inspection Data", IEEE Trans. Power systems, Volume 19, Issue 2, Page(s): 782 – 787.

[6]- Radmer, D.T., Kuntz, P.A., Christie, R.D., Venkata, S.S.; Fletcher, R.H., 2002, "Predicting Vegetation-Related Failure Rates for Overhead Distribution Feeders", IEEE Trans. Power Delivery, Volume 17, Issue 4, pp. 1170 – 1175.

[7]- Xu, L., Mo-yuen Chow, Taylor, L., S., 2003, "Analysis of Tree-Caused Faultsin Power Distribution Systems", in Proc. North Amer. Power Symp.

[8]- Zhou, Y., Pahwa, A., 2007, "Modeling Weather-Related Failures of Overhead Distribution Lines", in *Proc*. IEEE PES General meeting, pp1 -1.

[9]- Cochenour, G., Das, S., Pahw, A., 2008, "A Multi-Objective Evolutionary Strategy Based Radial Basis Function Network Approach for Predicting Failure Rates in Distribution Systems", ISAST Trans. on Intelligent Systems, No. 1, Vol. 1, pp. 7-14.

[10]- R. M. Bucci, R. V. Rebbapragada, A. J. McElroy, E. A. Chebli, and S. Driller, 1994, "Failure predic-tion of underground distribution feeder cables," IEEE Trans. Power Delivery, vol. 9, pp. 1943–1955.

[11]- Pylvanainen, J., Jarvinen, J., Verho, P., 2004, "Advanced Reliability Analysis for Distribution Network", IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies, Volume 2, pp. 457 - 462 Vol.2.

[12]- Falaghi, H., Haghifam, M. –R., Tabrizi, M., 2005, "Fault Indicators Effects on Distribution Reliability Indices", Proceedings of the CIRED 2005, 6-9 June, Turin, Italy.

[13]- The Greater Tehran Electricity Distribution Company portal: http://www.gtedc.ir, retrieved at: 04.01.2009.

[14]- Chow, M.-y., Yee, S.O., Taylor, L.S., 1993 "Recognizing Animal-Caused Faults in Power Distribution Systems Using Artificial Neural Networks", IEEE Trans. Power Delivery, Volume 8, Issue 3, pp.1268 – 1274

[15]- Fereidunian, A.R., Lesani, H., Lucas,C., 2002, "Distribution System Reconfiguration Using Pattern Recognizer Neural Networks" International Journal of Engineering (IJE), Transaction B: Applications, Vol.15, No.2, pp.135-144.

[16]- Xu, L., Mo-Yuen Chow, 2006, "A Classification Approach for Power Distribution Systems Fault Cause Identification", IEEE Trans. Power systems, Volume: 21, Issue 1, pp. 53- 60.

[17]- Bishop, C.M., 2004, "Neural Networks for Pattern Recognition", Oxford University Press.

X. Biographies

Elham Akhavan-Rezai received her BS from Guilan University, Rasht, Iran in 2005; and she is pursuing her MSc Thesis on power system reliability modeling at the Islamic Azad University, Tehran-South Branch, in cooperation with the GTEDC. Her research interests include power distribution system reliability, preventive maintenance, and distribution automation. She is a student member of IEEE.

Mahmood-Reza Haghifam was born in Iran in 1967. He received the BS, MSc and PhD degrees in electrical engineering in 1989, 1992 and 1995. He is a Professor in Power Systems at the Tarbiat Modarres University (TMU). He is a Senior Member of the IEEE and Research Fellow of Alexander von Humboldt in Germany. He has been awarded research grants from AvH, DAAD. Also he was a visiting professor in university of Calgary, Canada in 2004. His main research interests are electric distribution systems, power system reliability, power system restructuring and soft computing applications in power system analysis and operation.

Alireza Fereidunian earned his PhD and MSc degrees from University of Tehran in 2009 and 1997, and BSc from IUST in 1993. He is currently a research associate with the CIPCE, University of Tehran; and an independent consultant to the GTEDC. He has been awarded research grants from The British Council (Imperial College London), CIMO (Helsinki University of Technology) and IDB Merit Scholarship. His research interests include automation and operation of distribution systems, AI and expert systems in power systems, systems engineering, decision support systems, and human-automation interactions. He is a member of IEEE and INCOSE.