Faulted zone determination using statistical modeling of voltage sag database in power distribution systems

Determinación de la zona en falla usando modelado estadístico de bases de datos de huecos de tensión en sistemas de distribución de energía eléctrica

Gabriel Ordóñez Plata[1], Jorge Cormane Angarita[1], Juan Mora Flórez[2]*

[1]Grupo de Investigación en Sistema Eléctricos (GISEL). Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Universidad Industrial de Santander. Calle 9, Cra 27 Ciudad Universitaria. Bucaramanga, Colombia

[2]Grupo de Investigación en Calidad de Energía Eléctrica y Estabilidad (ICE[3]). Programa de Ingeniería Eléctrica. Universidad Tecnológica de Pereira. La Julita, Ciudad Universitaria. Eléctrica 227. Pereira, Risaralda, Colombia

(Recibido el 14 de marzo de 2008. Aceptado el 6 de noviembre de 2008)

Abstract

An alternative solution to the problem of power service continuity associated to fault location is presented in this paper, by using a methodology of statistical nature based on finite mixtures. A statistical model which helps to locate the faulted zone, is obtained from the extraction of the magnitude of the voltage sag registered during a fault event, along with the network parameters and topology. The objective is to offer an economic alternative of easy implementation for the development of strategies oriented to improve the reliability from the reduction of the restoration times in power distribution systems. As results presented for an application example in a 25kV system, the faulted zones were identified, having low error rates.

----- *Keywords*: Power Quality, Fault Location, Finite Mixtures, Statistical Model, density mixture models.

Resumen

En este artículo, se presenta una solución alternativa para el problema de continuidad del servicio asociada a la localización de fallas. La metodología

[*] Autor de correspondencia: teléfono: + 57 + 6 + 334 40 00, fax: + 57 + 6 + 335 96 21, correo electrónico: gaby@uis.edu.co (G. Ordóñez).

propuesta es de naturaleza estadística y basada en las mezclas finitas. El modelo estadístico es obtenido a partir de la extracción de la magnitud del hueco de tensión registrado durante un evento de falla y de los parámetros de la red y de su topología. El objetivo esta asociado a ofrecer una alternativa económica y de fácil implementación para el desarrollo de estrategias orientadas a mejorar la confiabilidad a partir de la reducción de los tiempos de restauración de los sistemas de distribución. Como resultados más importantes, se presentan los obtenidos en un ejemplo de aplicación en un sistema de 25 kV, en el cual las zonas en falla fueron localizadas con un bajo error en el desempeño del localizador.

----- *Palabras clave*: Calidad de potencia, localización de fallas, mezclas finitas, modelos estadísticos, modelos de mezclas de densidad.

Introduction

The interest in improving quality of supplied power is due to the deregulation in the electrical industry where quality is not only an indicator of the participation in the open power market but also one of the most relevant aspects regarding the requirements imposed to utilities. In most of the countries, and as a consequence of the new regulation, it is intended to strengthen the business of electricity distribution and the market from the viewpoint of power quality [1, 2]. The dependency of human activities in electricity demands that energy be supplied under several criteria of security, reliability and quality [3]. Such criteria have been promoted by fixed prices charged to final customers and by standards imposed to the service provided by utilities [4]. The continuity of supply is one of the most important aspects for the customer. This importance emerges from the social and economic impacts of interruptions [5, 6]. Although it is not economically feasible to reach a 100% of reliability, utilities are making efforts to mitigate the problem of interruptions with an adequate planning and operation of the power system [7].

According to the statistics data, about 80% of interruptions are caused by faults in the distribution system. The application of transmission system fault location algorithms to distribution networks is not a easy task due to the topology and operating principles of the latter (i.e. non homogeneous feeders, load taps, laterals, radial operation and the available measuring equipment) [8]. There exists a variety of methods for locating faults in power distribution systems. These methods may be classified in three broad categories. The first one comprises methods that detect components of high frequency in travelling waves, the second includes methods that compute fault impedance from the rms values of current and voltages measured at the fundamental frequency, and the last one is based on methods of visual inspection that consist of patrolling and checking the faulted feeder [8, 9].

This paper is aimed to propose an alternative solution to the problems associated with interruptions by means of a statistical model of voltage sags database applied to determine the fault location in power distribution systems to reduce the time wasted in system restoration [10]. The achievement of this goal enables the improvement of reliability from the establishment of strategies which are both economic and easily applicable by utilities [11].

This paper consists of six sections. In section 2, the theory related to the method which the proposed approach is based and the requirements for obtaining the statistical model are introduced. In section 3, the methodology proposed to fault location in distribution systems is presented. Section 4 describes the algorithm. Section 5 presents an application example by using a 25 kV distribution system, where the results obtained with the application of the proposed methodology are shown. Finally, the conclusions are highlighted in section 6.

Basics of the multivariate analysis

The multivariable analysis is used in the proposed approach as kernel to obtain a representation of the distribution system. Multivariate analysis is here decomposed in two main levels.

The first one consists of the extraction of information from available data, called Exploratory Data Analysis-EDA. In this case, the available data is composed by the fault registers of voltage measured at the power substation. These registers are characterized and as a result the voltage sag magnitude is obtained [12].

Second level is intended to represent knowledge from using the characteristics obtained at the first stage and relate it to the fault location.

Considering the above proposed, two different techniques are used to develop the fault location model. The first one is the application k-means algorithm and the second one is the mixture of distributions-MD [13]. With the application of the first technique data exploration and definition of variables are achieved while in the last one the probability density function is estimated.

Visual exploration is a powerful tool that serves as a first step in the understanding of multivariate data and enables the information analysis. This process helps with the comprehension despite the complexity and volume of data [14, 15].

Clustering data allows the conformation of meaningful groups in an analytical way, with the objective of classifying data in a population according to similarities or affinities [16]. Moreover, the use of relatively simple models for each local structure makes the implementation, analysis and computational simplification less difficult [17, 18]. The clustering algorithms of are based in the use of metric differences for the calculation of the distance. The metrics are subjected to the constraints in (1), where ***A***, ***B*** and ***C*** are individuals of a group of data and *d(**A**,**B**)* is the distance between individuals ***A*** and ***B***.

There exists a great variety of metrics associated to the quantification of data variability. In this approach the Euclidean (2) and Mahalanobis (3) metrics are used for cluster analysis since the best results have been achieved with them [19]. Equations 2 and 3 are written in matrix format, where $\mathbf{x}_i$ is the data vector which corresponds to observation *i*, $\bar{\mathbf{x}}$ is the mean magnitude vector, T indicates the transpose of a matrix and **V** is the covariance matrix.

$$\begin{aligned} &d(\mathbf{A},\mathbf{B}) \geq 0 \quad \forall \quad (\mathbf{A},\mathbf{B}) \\ &d(\mathbf{A},\mathbf{B}) = 0 \quad \Leftrightarrow \quad \mathbf{B} = \mathbf{A} \\ &d(\mathbf{A},\mathbf{B}) = d(\mathbf{B},\mathbf{A}) \\ &d(\mathbf{A},\mathbf{B}) \leq d(\mathbf{A},\mathbf{C}) + d(\mathbf{C},\mathbf{B}) \end{aligned} \tag{1}$$

$$dE_i = \left[(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \right]^{1/2} \tag{2}$$

$$dM_i = \left[(\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{V}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \right]^{1/2} \tag{3}$$

Two important aspects are evaluated by using the previous mentioned metrics. The first is concerned with the proximity of between elements of the same group which indicates the how compact is a group of data (Internal homogeneity). The second is related to the distance between elements of different groups in order to guarantee that no overlap exists due to the relative proximity of groups (external heterogeneity) [19].

MD is a technique of statistical modelling which allows an estimation of the Probability Density Function – PDF of data in a random sample, represented as a finite weighted sum of multivariate density components [20]. This technique has been applied with several purposes such as the modelling of heterogeneity in a population (biology), management of outliers, PDF estimation and clustering (statistics), pattern recognition (image processing) and fraud detection (utilities). Within the principal features of MD, it is worth to mention that a smoothing parameter for PDF estimation is not required. The finite number of terms in the mixture can be defined according to the needs of the analyst. MD have less computational burden in comparison with other methods such as Kernel Estimation and Histograms since less amount of information is stored for PDF estimation [19].

The main advantage of MD is due to the capability on analysis and modelling of clusters. The flexibility of mixture models allows its application to the fault location problem in distribution systems [21]. In this particular application, concepts about multivariate data were used. These concepts are a characterization of a multidimensional random phenomenon. The multivariate density mixtures are expressed as shown in (4).

$$f(\mathbf{x}) = \sum_{g=1}^{G} p_g f_g(\mathbf{x}, \boldsymbol{\theta}) \qquad (4)$$

MD taken from a random sample **x** of *n* observations of dimension *d* is comprised of *G* components f_g (x,θ) related to the selected multivariate density function. Each component f_g describes the behaviour of a group within the sample in which data related to such group have similar characteristics established in the estimation vector θ that corresponds to the parameters of each distribution (homogeneous). An estimator is a parameter that defines the behaviour of data in a group, so it also describes the shape. For example, estimators for a normal PDF are: the mean vector μ and the covariance matrix **V** which define the central point of the distribution and how data are concentrated. The quantities p_g, called weights or coefficients of mixture, provide information about the importance of the group within the mixture. The conditions that coefficients must satisfy are given in (5).

$$\sum_{g=1}^{G} p_g = 1 \qquad p_g > 0 \qquad (5)$$

The aim of MD is to identify an unknown quantity of groups in which data of a given population are clustered. That is to say, MD seeks the homogeneity within an initially heterogeneous sample. To achieve this goal MD utilizes the Expectation-Maximization-EM algorithm for parameter estimation.

The EM algorithm is an application of the Maximum Likelihood Method-MLM to find missing parameters. In this sense, it allows the determination of maximum likelihood estimators θ for each distribution f_g (x, θ)from initial values $\hat{\boldsymbol{\theta}}^{(\mathbf{0})}$[13]. *L* in the algorithm is the likelihood function of the sample, finding the expected value of the functions of missing values **Z** from the calculation of its density function with initial values $\hat{\boldsymbol{\theta}}^{(\mathbf{0})}$ and observed values **Y**. The result of this operation is called the E-step (Expectation) given in (6).

$$L(\boldsymbol{\theta} \mid \mathbf{Y}) = E\left[l(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z})\right] \qquad (6)$$

In the M-step (Maximization) the function $L(\boldsymbol{\theta} \mid \mathbf{Y})$ is maximized in θ to find the maximum likelihood estimators from the replacement of missing with estimated values.

Let $\hat{\boldsymbol{\theta}}^{(k+1)}$ be the value of the estimator obtained in the M-step. Then we return to the E-step in an iterative procedure until convergence is reached as presented in (7).

$$\left|\hat{\boldsymbol{\theta}}^{(k+1)} - \hat{\boldsymbol{\theta}}^{(k)}\right| \leq \varepsilon \qquad (7)$$

Where ε is the accepted tolerance, and *k* is the iteration number [22].

The MLM selects a good value of the estimator the one that maximizes the probability of generating the observed sample from the model to estimate [13, 21].

Proposed methodology

To obtain the fault location, a characterization of the system response under fault conditions is proposed. The response of the system is reflected in voltage signals measured at the distribution substation (single end measurements). In what follows, a description of the research done in five stages proposed methodology will be given.

In the first stage, a power distribution system is considered and voltage waveforms from faults are recorded at the substation. The most common type of faults in distribution systems are short circuits. This work takes into account single line-to-ground, line-to-line, double line-to-ground, three phase and three phase-to-ground faults with different values of fault resistance between 0 and 50 Ohms [23].

In the second stage, signals are pre-processed to obtain rms and per-unit values.

In the third stage the system was characterized in two different ways. The first is a deterministic one based on the calculation of single-phase features or "descriptors" of voltage signals [12]. The second of statistical nature is based on an application of EDA which results in descriptors that characterize system behaviour.

In the fourth stage the information of included descriptors is analyzed by setting rules and conditions for creating characteristic zones and relations of homogeneity between groups. This is achieved with the application of techniques for the analysis of clusters.

Finally, in the fifth stage, the model of mixtures is conformed and fault data are classified obtaining the most probable fault location zone.

Structure of the fault locator

The description of the algorithms that allow a step by step construction of the model, which is based on clustering and MD theory, is presented in this section. The idea for the implementation of the algorithm is to progressively adjust the model of each cluster extracted from the information of the system under fault conditions. The initial values of model parameters are calculated in an iterative procedure. A summary-type scheme of the algorithms known as k-means is next presented:

a. Specify the number of groups by a preliminary analysis or considering suggestions of the maintenance crew.

b. Determine the centres of these groups. This can be done a priori or in a random way.

c. Take each data and calculate the distance from each cluster by using Euclidean or Mahalanobis approaches as presented in equations (2) and (3).

d. Aggregate each data into the cluster whose distance is a minimum and compute the new centres.

e. Repeat steps b, c and d until no further changes occur in groups.

The idea is to minimize the sum of squared distance from points to centres within each group. Once the clustering is done, the method proceeds with the identification of predominant characteristics in each group with the aim of inferring and relating new data.

From the information about the groups from the k-means algorithm, initial values for the centres are calculated. The initial value of covariance matrix is taken as the identity matrix and the mixture coefficients are calculated with the proportion of data in each group, in relation to the sample. Once initial parameters have been obtained, the estimation of the mixture model parameters is initiated by the EM algorithm in an iterative procedure until desired convergence is reached. The results are the final values of parameters $\boldsymbol{\mu}$, **V** and p of each group. These parameters are used by the multivariable density of mixtures presented in equation (4), in order to classify other possible observations. The steps of the EM algorithm are as follows:

a. Determine the number of components of the mixture by using the k-mean algorithm.

b. Determine initial values of parameters of each component. ($\hat{\boldsymbol{\mu}}^{(0)}$, $\hat{\mathbf{V}}^{(0)}$ and $\hat{p}^{(0)}$).

c. Calculate the posterior probability for each observation (E-step) as shown in (8) and (9).

$$\hat{\tau}_{ij} = \frac{\hat{p}_i \phi\left(\mathbf{x}_j ; \hat{\boldsymbol{\mu}}_i , \hat{\mathbf{V}}_i\right)}{\hat{f}\left(\mathbf{x}_j\right)} \tag{8}$$

$$\hat{f}\left(\mathbf{x}_j\right) = \sum_{g=1}^{G} \hat{p}_g \phi\left(\mathbf{x}_j ; \hat{\boldsymbol{\mu}}_g , \hat{\mathbf{V}}_g\right) \tag{9}$$

$\hat{\tau}_{ij}$ represents the posterior probability of x$_j$ corresponding to the i term, $\phi\left(\mathbf{x}_j;\hat{\boldsymbol{\mu}}_i,\hat{\mathbf{V}}_i\right)$ is the normal multivariate density and $\hat{f}\left(\mathbf{x}_j\right)$ corresponds to the estimated MD for the i terms evaluated in x$_j$. j is a index which indicate the total amount of data.

d. Update $\hat{\boldsymbol{\mu}}$, $\hat{\mathbf{V}}$ and $\hat{p}$ of each component (M-step) by using (10), (11) (12). ($\hat{p}_i$, $\hat{\boldsymbol{\mu}}$ y $\hat{\mathbf{V}}_i$ are the updated estimations of the parameters).

$$\hat{p}_i = \frac{1}{n}\sum_{j=1}^{n}\hat{\tau}_{ij} \tag{10}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n}\sum_{j=1}^{n}\frac{\hat{\tau}_{ij}\mathbf{x}_j}{\hat{p}_i} \tag{11}$$

$$\hat{\mathbf{V}}_i = \frac{1}{n}\sum_{j=1}^{n}\frac{\hat{\tau}_{ij}\left(\mathbf{x}_j-\hat{\boldsymbol{\mu}}_i\right)\left(\mathbf{x}_j-\hat{\boldsymbol{\mu}}_i\right)^T}{\hat{p}_i} \tag{12}$$

e. Repeat steps c and d until desired convergence is obtained.

Subsequently, the organization of groups in classes associated to faults is based in the probability of appearance in each group as given by the mixture model (13).

$$\hat{f}_{FM}(\mathbf{x}) = \sum_{g=1}^{G} p_g\phi_g(\mathbf{x};\hat{\boldsymbol{\mu}}_g,\hat{\mathbf{V}}_g) \tag{13}$$

Proposed tests for locating faults

From the statistical model as presented before, fault location within the system according to its response is expected. Recorded voltage waveforms are the basis to find the solution. Each recorded event has relevant information that enables data classification within certain type of class established in the model. Each class corresponds to a zone within the distribution network.

All the information used in this approach corresponds to magnitudes of voltage (not angles were used). That is because this approach is aimed to obtain a low economical cost tool due to constrains imposed in several distribution utilities.

A 25 kV power distribution system is proposed for tests. This system is taken from Saskatown Power and Light, Canada, and it is presented in figure 1 [8, 24].

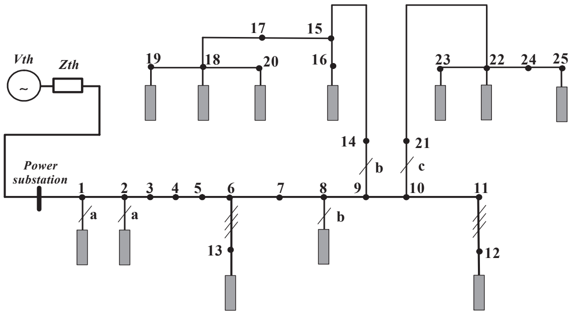

Figure 1 25 kV power distribution system used to test

For the discrimination of zone visually, detectable groups were taken into account in a preliminary data analysis. Also, the zone division of the power system has to consider the suggestions of the maintenance crew, according to their experience in fault recovering. The goal to be achieved by associating groups to zones is to establish the correspondence between fault location and data classification within groups. Three descriptors were used to represent distribution system information: maximum sag magnitude in each phase of a three phase system [12].

In this case, data for training and validation are short-circuit faults simulated in each bus of the power distribution system by using different fault resistances from 0,05Ω to 50Ω according to [23]. In table 1, fault resistance values used in the training (T) and validating (V) processes are presented.

In table 2 data used in training (T) and validation (V), for each fault type are presented.

Before classification process, a verification of training data is performed to previously identify and divide the distribution network in several zones. The objective is to estimate the initial centres of each group. Figure 2 shows each of the power system zones obtained previous grouping of voltage sag descriptors for single-phase faults.

Table 1 Fault resistance values

Fault resistance [Ω]	*Process*	*Fault resistance [Ω]*	*Process*
0.05	T	25	T
5	T	30	V
10	V	35	T
15	T	40	V
20	V	50	T

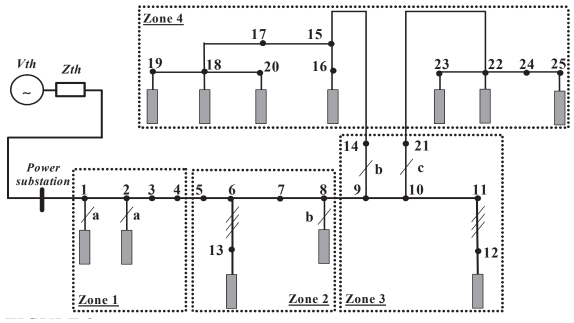

Figure 2 Zones previously determined by visual analysis of data

Table 2 Data used in training and validating processes

Fault type	*Simulations by fault type*			*Process*	
				Training (T)	*Validation (V)*
	A	*B*	*C*		
Line to ground	132	187	176	315	180
	AB	BC	CA		
Line to line	132	132	132	252	144
Double line to ground	132	132	132	252	144
		ABC			
Three phase		132		84	48
Three phase to ground		132		84	48
Total		1551		987	564

The proposed model is ordered in several steps of classification, the first one (SC1), determines the faulted phase; the second (SC2), finds fault resistance value and the third (SC3), finds fault location. SC1 step is only applicable to single-phase and phase-to-phase faults while SC2 and SC3 steps are applicable to all fault types. Figures 3, 4 and 5 show the distribution of training data corresponding to three types of faults. Besides, three clearly defined groups which correspond to each one of the three phases are observed.

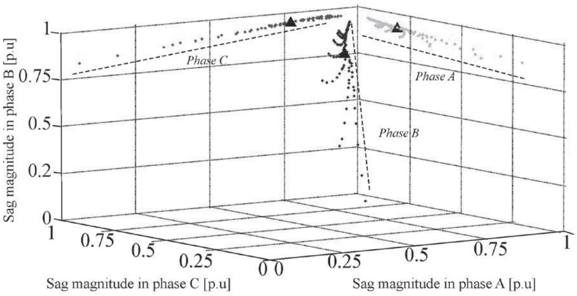

Figure 3 Distribution of training sag data for single-phase fault (SC1)

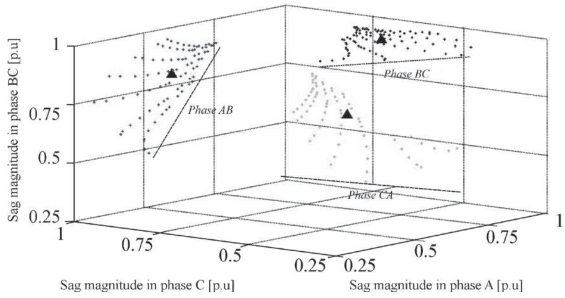

Figure 4 Distribution of training sag data for phase-to-phase fault (SC1)

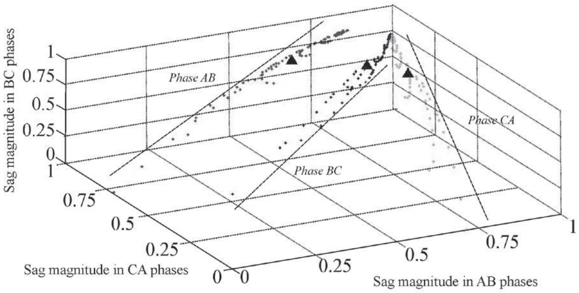

Figure 5 Distribution of training sag data for double-phase-to-ground fault (SC1)

Having the initial definition of groups and by applying the k-means algorithm, centers of groups are estimated to allow initialization of the method. Then, the shape and final proportion of groups within the distribution is defined utilizing EM algorithm and the initial estimation. In this step, a heterocedastic model is used to determine the shape of the covariance matrices for each distribution and also the shape of each group. Moreover, by using the same initial values for mixture coefficients of each group, it assumes that the occurrence of a fault within the group is equally probable.

In step SC2, the information about fault resistance allows the creation of possible scenarios where groups representing intervals of fault resistance values are conformed. In this particular case, five representative groups were established, as shown in figure 6. The idea presented above is related to the capability of predicting what caused the fault. This information can be relevant to establishing the procedure to solve the problem [23].

The shape and final size of the five groups obtained at this application is shown in figure 6 (ellipses). Centres and mixture coefficients are represented by stars. Fault resistance value grows advancing to the right in the figure. The elliptic shape of each group is related to the values of the covariance matrix [16]. At step SC3 all data are grouped, and each one represents a probable faulted zone.

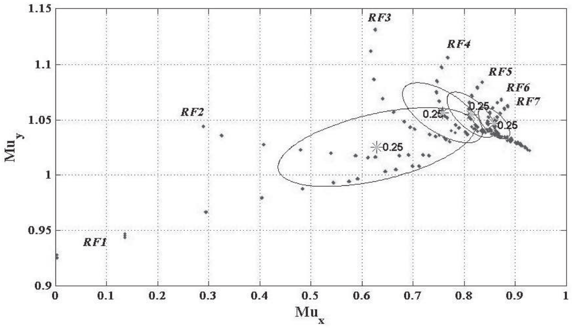

Figure 6 Distribution of groups in training sag data in case of single-phase faults (SC2)

In SC3 step, data are clustered and used as zones of fault occurrence. Each defined zone is associated to a given number of buses within the distribution network as shown in figure 2. These zones are established with the information about the number of groups obtained in SC2 step. Hence in step SC3 we have "*r*" groups corresponding to fault resistance groups determined in SC2 and these contain "*z*" groups corresponding to each one of system zones. In the case of the prototype system, there is a MD that consists of five groups (SC2) and each group contains four subgroups that represent system zones. Figure 7 illustrates the construction of groups associated to zones.

The EM algorithm is able to modify the elements of the covariance matrix in each one of the iterations until the best estimation is obtained. If the covariance matrices remained without change from the initial estimation, circular shaped groups would be obtained [19]. Finally, the mixture model is applied with the aim of comparing initially assumed zones (see figure 2), with those determined by the algorithm. In the current application case, the initial estimate of zones corresponds to the result obtained with the algorithms. An aspect to remark is the incidence of the initial values, assumed by the analyst, in the improvement of the model during its construction. Once established the model, validation was done with sag data not used in the training process.

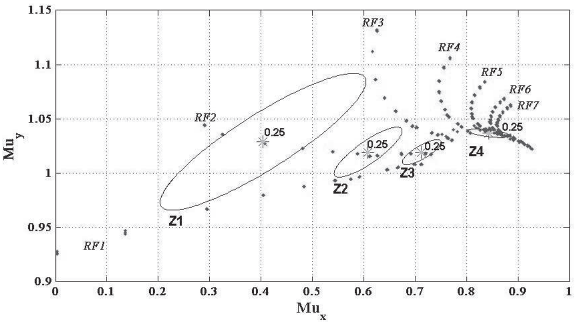

Figure 7 Distribution of zones associated to group 1 in SC2 step

Table 3 Results of SC1 Validation Step

Fault type	*Phase (s)*			*Subtotals*
	A	*B*	*C*	
Line to ground	48/48 : 100	68/68 :100	64/64 : 100	180/180 : 100
	AB	BC	CA	
Line to line	48/48 : 100	48/48 : 100	48/48 : 100	144/144 : 100
Double line to ground	39/48 : 81.25	37/48 : 79.16	37/48 : 77.08	114/144 : 79.16

The observations contain voltage sags with fault resistance values of 10, 20, 30 and 40 Ω. Having established the data model, validation stage is performed by using data not previously used in training. The validation results are presented as the ratio between the well classified data and all the testing data. To evaluate the fist classification stage (SC1), 468 faults were used. In table 2 all data used in validation are presented. In table 3 the results in case of single and double phase faults are presented.

In Table 4, validation results obtained with SC2 classifier are presented (564 observations).

The results in table 5 reflect the good performance of SC3 classifier (564 observations).

Classification problems are found in further points along the distribution feeder. In these cases there exists a wrong assignation of the zone due to a classification error in the previous step SC2. Another possible cause of this inadequate behaviour is the location of data in the intersection of two or more groups. In this situation, the zone is assigned depending on its higher probability *a posteriori.*

Each observation is evaluated respect to each one of the groups by computing the probability of pertaining to the same groups through all the steps. The observation is classified within the group in which it has the higher probability.

Table 4 Results of SC2 Validation Step

Fault type	*Fault resistant range*				*Subtotals*
	R1	*R2*	*R3*	*R4*	
Line to ground	27/45 : 60	36/45 : 80	36/45 : 80	45/45 : 100	144/180 : 80
Line to line	27/36 : 75	30/36 : 83.33	36/36 : 100	36/36 : 100	129/144 : 89.58
Double line to ground	28/36 : 77.77	30/36 : 83.33	36/36 : 100	36/36 : 100	130/144 : 79.16
Three phase	20/24 : 83.33	24/24 : 100	24/24 : 100	24/24 : 100	90/96 : 93.75

Table 5 Validation results in stage SC3

Fault type	*Zones*				*Subtotals*
	Z1	**Z2**	**Z3**	**Z4**	
Line to ground	48/48 : 100	48/48 : 100	48/48 : 100	28/36 : 77.77	172/180 : 95.55
Line to line	48/48 : 100	48/48 : 100	48/48 : 100	-	144/144 : 100
Double line to ground	34/48 : 70.08	43/48 : 89.58	36/48 : 75.00	-	113/144 : 78.87
Three phase	32/32 : 100	32/32 : 100	32/32 : 100	-	96/96 : 100

Table 6 presents results of data classification according to calculated probabilities in each group. Eight data are presented as an example due to the great amount of processed information (468 observations). Data shown corresponds to eight single-phase faults located in the four zones of the system in figure 2. Each fault belongs to a real zone of occurrence and presents an estimated zone which is determined from its probability of occurrence calculated in the classification performed in stage SC3. No information is included about three phase and phase-to-phase faults because only the main three phase feeder is shown and divided in three zones. On the other hand, single phase feeders pertaining to the fourth zone of the system are not included.

Conclusions

The proposed statistic based methodology for fault location in distribution systems has been presented and tested. This approach is based in the statistical modelling and extraction of the sag magnitude from voltage measurements stored in fault data bases. The fault locator here proposed contributes to satisfy actual needs of utilities in preserving and improving service quality, promoting the consolidation of strategies oriented to decrease the number and duration of interruptions. The total time of interruption can be associated to the time taken in several actions during system restoration such as alarm time, detection time, access time and sectioning time among others.

Potential limitations for the proposed methodology are the selection of the number of groups, the proportion of samples in each group and initial values required by algorithms. These difficulties overcome by introducing theoretical and heuristic criteria, the latter being more influential in the structure of the model.

One of the advantages in the construction of the model is the determination of groups of well defined characteristics which allow an optimization in the classification of data thus ensuring good model accuracy. Moreover, the initial approximation of system zones is useful in the estimation of the real number and size of zones in the distribution system. Besides, detailed characterization of the system, model performance depends on the quality of information extraction and processing.

Table 6 Examples of results of data classification

Data	*Real Zone*	*Probability in each group [%]*				*Estimated Zone*
		Z_1	Z_2	Z_3	Z_4	
1	1	98.98	0.42	0.35	0.25	1
2	1	99.88	0.08	0.03	0.01	1
3	2	0.99	95.4	1.98	1.63	2
4	2	2.89	94.94	1.62	0.55	2
5	3	2.92	0.61	82.35	14.1	3
6	3	0.79	8.29	82.88	8.04	3
7	4	2.35	5.28	38.27	54.51	4
8	4	0.26	15.42	25.85	58.47	4

Software requirements for the implementation of a tool with the proposed methodology are supplied by statistical packages of commercial use and low license costs. This fact enables the easy and simple implementation of the model. Hardware requirements are basic and economical since they only depend of a data acquisition system with the capability of continuously monitoring power system magnitudes at the power distribution substation.

The proposed model takes into account technical, economical and operational issues existing in power distribution networks. A remarkable aspect is the low investment cost for the implementation of the fault detection system based in the proposed method.

References

1. J. Mora, G. Carrillo, B. Barrera. "Fault Location in Power distribution Systems using a Learning Algorithm for Multivariable Data Analysis". *IEEE Trans. on Power Delivery*. Vol. 22. 2007. pp. 1715-1721.
2. J. Driesen, T. Green, T. Van Craenenbroeck, R. Belmans. "The Development of Power Quality Markets". *IEEE Power Engineering Society General Meeting.* Vol. 1. 2004. pp. 963-967.
3. J. Martinez, J. Martin. "Voltage Sag Stochastic Prediction Using an Electromagnetic Transients Program". *IEEE Transactions on power delivery.* Vol. 19. 2004. pp. 596-602.
4. C. Crozier, W. Wisdom. "A power quality and reliability index based on customer interruption costs". *Power Engineering Review, IEEE*. Vol. 19. 1999. pp. 59 – 61.
5. M. Bollen. *Understanding power quality problems: voltages sags and interruptions*. Ed. IEEE Press. New York,. 2000. pp. 35 – 116.
6. A. Girgis, C. Fallon, D. Lubkeman. "A fault location technique for rural distribution feeders". *IEEE Trans. Industry applications.* Vol. 29. 1993. pp. 1170-1175.
7. R. Brown. *Electric power distribution reliability*. New York. Marcel Dekker. 2002. pp. 24-42.
8. R. Das. "*Determining the Locations of Faults in Distribution Systems*". Ph.D dissertation. Saskatchewan Univ. Canada. 1998. pp. 15 – 48.

9. J. Zhu, D. Lubkeman, A. Girgis. “Automated fault location and diagnosis on electric power distribution feeders”. *IEEE Trans. Power delivery.* Vol. 12. 1997. pp. 801-809.

10. A. Girgis, C. Fallon, D. Lubkeman. “A fault location technique for rural distribution feeders”. *IEEE Trans. Industry applications.* Vol. 29. 1993. pp. 1170-1175.

11. H. Willis. *Power distribution planning reference book*. New York. Marcel Dekker. 2004. pp. 47 – 59.

12. L. D. Zhang, M. H. J. Bollen. “Characteristics of voltage dips (sags) in power systems”. *IEEE Transactions on Power Delivery*. Vol. 15. 2000. pp.827-832.

13. A. Rencher. “*Methods of Multivariable Analysis*”. Ed. John Wiley and Sons. New York. Brigham Young University. Utath. 1995. Chapter 12. pp. 415-443.

14. Y. Wang, L. Luo, M. Freedman, S. Kung. “Probabilistic principal component subspaces: A hierarchical finite mixture model for data visualization”. *IEEE Trans. Neural Networks.* Vol. 11. 2000. pp. 625-636.

15. R. Johnson, D. Wichern. *Applied Multivariate Statistical Analysis*. Ed. Prentice Hall, New York. 1998. pp. 124-168.

16. J. Hair, R. Anderson, R. Tatham, W. Black. *Multivariable Data Analysis*. Ed. Prentice Hall. Madrid.1999. pp. 85-97.

17. Y. Wang, S. Lin, H. Li, S. Kung. “Data mapping by probabilistic modular networks and information theoretic criteria”. *IEEE Trans. Signal processing.* Vol. 46. 1998. pp. 3378-3397.

18. M. Jordan, R. Jacobs. “Hierarchical mixture of experts and the EM algorithm”. *IEEE Trans. Neural Computing*. Vol. 6. 1994. pp. 181-214.

19. W. Martínez, A. Martínez. *Computational statistics Handbook whit MatLab*. Ed. Chapman & Hall New York. 2002. pp. 90-124.

20. E. Dalla Jonson. *Métodos multivariados aplicados al análisis de datos*. Ed. Thompson. México. 2000. pp. 53-86.

21. G. Mclachlan, D. Peel. *Finite mixture models.* Ed. Wiley and Sons. Montreal. 2000. pp. 46-72.

22 M. Figueiredo. “Unsupervised Learning of Finite Mixture Models”. *IEEE Trans. Pattern analysis and Machine intelligence*. Vol. 24. 2002. pp. 135-143.

23. J. B. Dagenhart. “The 40-Ground-Fault Phenomenon”. *IEEE Trans. on Industry Applications*. Vol. 36. 2000. pp. 30-32.

24. S. Lee, M. Choi, S. Kang, B. Jin, D. Lee, B. Ahn, N. Yoon, H. Kim, S. Wee. “An intelligent and efficient fault location and diagnosis scheme for radial distribution systems”. *IEEE Trans. Power Delivery.* Vol. 19. 2004. pp. 524-532.