Fault zone classification in power distribution systems using clustering algorithms

Clasificación de zonas en fallas en sistemas de distribución de energía utilizando algoritmos de agrupamiento

DOI: https://dx.doi.org/10.17981/ingecuc.21.1.2025.09

Original Research.
Date Received: 08/08/2023, Date Accepted: 06/09/2023.

Daniela Carolina Castillo-Acosta
Universidad del Magdalena. Santa Marta, Colombia
danielacastilloca@unimagdalena.edu.co

Carlos Robles-Algarín
Universidad del Magdalena. Santa Marta, Colombia
croblesa@unimagdalena.edu.co

Luis Camargo-Ariza
Universidad del Magdalena. Santa Marta, Colombia
lcamargoa@unimagdalena.edu.co

To cite this paper
D. Castillo-Acosta, C. Robles-Algarín & L. Camargo-Ariza "Fault zone classification in power distribution systems using clustering algorithms," *INGE CUC*, vol. 21, no. 1, 2025. DOI: https://dx.doi.org/10.17981/ingecuc.21.1.2025.09

Abstract

Introduction: The main cause of disruptions in the energy service is generated in the distribution grids. The faults need repairs before restoring the service, which requires their rapid location, since this process influences the duration and frequency of the disruptions. Many companies assess power service quality solely by measuring the equivalent interruption frequency and duration (DES and FES), as they lack the resources to invest time and money in strategies to improve system reliability.
Objective: Determine the location of the fault zone in medium voltage power distribution systems.
Method: The clustering methods of k-means and the Gaussian Mixture Model (GMM) based on the expectation-maximization algorithm were implemented, to create groups based on their characteristics, thus defining the probability of belonging to each one.
Results: A methodology for the detection of failures efficiently was obtained, which serves as support for planning processes and execution of action plans, facilitating the taking of corrective measures related to the continuity of the service and decreasing the system restoration time.
Conclusions: The application of the k-means and GMM algorithms allows identifying possible fault zones according to the test data. Although it does not show a single fault zone, since it makes estimates according to the data, it is a tool to make decisions based on the estimates obtained.

Keywords: k-means; power distribution systems; Gaussian mixture model; expectation-maximization algorithm; fault zones; clustering algorithms.

Resumen

Introducción: La mayor causa de las interrupciones del servicio de energía se genera en las redes de distribución. Las fallas requieren de reparaciones antes de restablecer el servicio, lo cual exige de su localización rápida, puesto que este proceso influye en la duración y frecuencia de las interrupciones. Muchas empresas solo realizan estudios de la calidad del servicio de energía a través de la medición de la frecuencia equivalente de las interrupciones y la duración equivalente de las interrupciones del servicio, debido a que no tienen la capacidad de invertir tiempo y dinero en estrategias para mejorar la confiabilidad del sistema.
Objetivo: Determinar la ubicación de la zona en falla en sistemas de distribución de energía de media tensión.
Metodología: Se aplicaron los métodos de agrupamiento k-means y el modelo de mezcla gaussiana basado en el algoritmo de maximización de expectativas, para la creación de grupos a partir de sus características, definiendo así la probabilidad de pertenencia a cada uno.
Resultados: Se estableció una metodología para la detección de fallas de forma eficiente, que sirve de soporte para procesos de planificación y ejecución de planes de acción, facilitando la toma de medidas correctivas relacionadas con la continuidad del servicio y disminuyendo el tiempo de restauración del sistema.
Conclusiones: La aplicación de los algoritmos k-means y modelo de mezclas gaussianas permite identificar posibles zonas en falla de acuerdo con los datos de las pruebas. Aunque no mostrará una zona única que se encuentre en falla, puesto que realiza estimaciones de acuerdo con los datos, es una herramienta para tomar decisiones según las estimaciones obtenidas.

Palabras clave: k-means; sistemas de distribución de energía; modelo de mezcla gaussiana; algoritmo de maximización de expectativas; zonas en falla; algoritmos de agrupamiento.

I. INTRODUCTION

The timely identification and resolution of failures in power distribution systems are a priority for energy service providers. This is not only due to their impact on quality indicators but also because of customer satisfaction, as end users are affected by service continuity interruptions. These aspects are crucial, as they contribute to ensuring the availability of an efficient energy supply, encouraging the implementation of measures to prevent disruptions, and promoting market liberalization to support free competition. Additionally, these factors facilitate the establishment of criteria for approving energy sales agreements between electric utilities and large consumers [1].

The Energy and Gas Regulatory Commission (CREG) and the Superintendence of Public Utilities (SSPD) have established energy service quality indicators to monitor and enforce compliance with quality standards. Through Resolution 070 of 1998, CREG introduced the DES (Equivalent Duration of Service Interruptions) and FES (Equivalent Frequency of Service Interruptions) indicators, which were the first quality metrics related to the number of interruptions occurring in a circuit and the average duration of service outages per customer. Moreover, fault detection and diagnosis facilitate faster system restoration, thereby reducing service discontinuity times. The determination of these indices enables the establishment of limits for network operators and customers, who, if non-compliant, are required to provide compensation [2].

In 2016, the SSPD conducted an analysis of the DES and FES indicators for 19 distribution companies covering 97% of the national demand. The study revealed significant regional differences across Colombia, with 40% of energy service providers exceeding the average DES. According to the DES indicator, 31.21% of the total duration is attributed to the departments of the Caribbean Region, Cundinamarca, Nariño, Caquetá, Tolima, and Chocó, where users experience an average of 74 cumulative hours of outages per year. This means that customers within this percentage endure approximately three days without electricity annually. Regarding the FES indicator, areas with values equal to or below 20 hours include Bogotá and parts of Valle del Cauca, covering 21.5% of the national demand. Santander, Norte de Santander, Antioquia, Caldas, Risaralda, Quindío, Cauca, and certain areas of Valle del Cauca fall within the 21 to 50-hour range, accounting for 42.2%. The remaining departments report an FES exceeding 50 hours, covering 33.1% of the national demand [3].

In 2019, a report on the quality of electricity service in Colombia, based on an assessment conducted by the SSPD in 2018, was presented. The report highlighted that ELECTRICARIBE S.A. E.S.P. served 18% of the country's users, representing approximately 2.3 million people. These users recorded the lowest service continuity indicator nationwide, with the poorest performance in every quality group analyzed compared to other Network Operators (OR). Furthermore, users in quality groups 2 and 3, representing around 590,000 customers, experienced a System Average Interruption Duration Index (SAIDI) nearly twice the national average. The company reported a SAIDI exceeding 108.2 hours, equivalent to 4 days and 12 hours without electricity, marking a slight improvement from the 112.8 hours recorded in 2017. Regarding the System Average Interruption Frequency Index (SAIFI), ELECTRICARIBE registered a value of 104.8 interruptions per user, more than double the national average of 48 interruptions per year [4].

The aforementioned situation is alarming regarding the quality of the service provided, as a reliable electricity supply is essential for most activities, including daily, domestic, and industrial operations. Moreover, electronic and electrical devices are increasingly susceptible to variations in power supply factors, which must meet specific reliability criteria [5].

Network Operators submit reports to the Unique Information System (SUI), which are used to calculate quality indicators. Service interruptions are recorded in databases of power distribution systems, making them a crucial subject of study for identifying failure characteristics through algorithmic analysis. The grouping of data sets is particularly relevant, as it enables a better understanding of the behavior of a population when only a sample of its components is available.

In power distribution systems, three-phase, two-phase, and single-phase connections can be found. Most medium-voltage networks are three-phase, although two-phase configurations are sometimes present in rural areas. However, at low-voltage levels, various types of

connections exist, with most residential loads being single-phase. Another key characteristic of these systems is the presence of diverse load types, including residential, commercial, industrial, and agro-industrial loads. Each of these has a typical power factor and specific behavior influenced by voltage and temperature variations. The most common causes of power losses include Joule losses, conductor and distribution transformer winding heating, hysteresis losses, eddy currents, and, notably, short-circuit faults [6]. For this reason, this study simulated different types of short-circuit faults, including single-phase, two-phase, and three-phase faults.

Among the alternatives used to prevent a power outage or blackout is load reduction, which involves voluntarily decreasing the supply voltage for a duration ranging from minutes to hours. These situations occur when the system is heavily loaded and lacks sufficient reserves to meet the energy demand. Such high demands can arise due to an abnormal load increase, often triggered by extreme weather conditions or the loss of a critical component, such as a transformer or a power line, due to a fault [7].

Techniques used for fault location in distribution systems must consider various characteristics, such as radial topology, extensive network branching, the presence of intermediate loads, and conductors with different gauges, among others. Traditional methods, such as impedance-based approaches, perform measurements at a single terminal. However, in highly branched networks, these methods can result in multiple possible fault locations, as the equivalent impedance may be the same at different nodes. Consequently, while these methods can estimate the distance to the fault, its exact location remains uncertain due to identical equivalent impedances in multiple branches. Additionally, these methods require highly accurate models of line and load characteristics [8].

An alternative approach to understanding fault behavior is through the development of clustering algorithms. Among parametric clustering methods, distribution mixture models stand out, as they allow for the modeling of probability density functions of datasets. In model-based clustering, mixture models with multivariate components are commonly used, often refined through probability-maximizing processes, such as the Expectation-Maximization (EM) algorithm [9]. With advancements in Monte Carlo sampling techniques, mixture models have gained popularity across various fields, including machine learning, biomedicine, and pattern recognition, among others [10].

The EM algorithm requires initial variables, which can be obtained from clustering algorithms such as k-means and initialized using the values obtained [11]. Among traditional clustering algorithms, such as k-means and spectral clustering, prior knowledge of the number of clusters is required, which can impact clustering performance [12]. Most clustering algorithms rely on assumptions to define the subgroups within a dataset. Consequently, the representation of clusters requires validation processes, which must address complex challenges such as cluster quality, the optimal number of clusters, and the level at which a dataset is established [13].

Among the tools used to address these challenges is the Calinski-Harabasz index, which helps determine the optimal number of clusters and improves the efficiency of the methodology. This cluster validity index is calculated for different data partitions, with the optimal partition corresponding to the highest index value [14].

Consequently, this research developed a methodology to determine the most probable fault location in a power distribution system by applying the k-means algorithm and the Gaussian Mixture Model (GMM) based on the EM algorithm. This approach serves as an effective tool, providing an accessible and easily implementable alternative that can help formulate strategies aimed at enhancing reliability and reducing restoration times in power distribution systems.

II. LITERATURE REVIEW

This section is divided into two parts: the first presents the state of the art on fault location methods in power distribution systems, and the second discusses the application of clustering algorithms in various fields.

A. Fault location methods

Among the methods used for fault location are impedance-based methods and other approaches based on fundamental frequency components. These methods are employed to calculate the fault distance from the primary distribution bus, which is estimated through an impedance-dependent method. To apply this approach, voltage and current values measured at the line terminals are required [15].

In [16], mathematical models were utilized to determine impedance values up to the fault point, enabling the estimation of fault distance for any distribution line. To validate this approach, a 132/33 kV secondary transmission substation was implemented in Nigeria, equipped with two 132/33 kVA, 60 MVA power transformers. Fault currents were extracted, and the maximum trip time required for protection devices was defined, facilitating the selection of the optimal relay and circuit breaker to ensure effective system operation before faults occur.

Using MATLAB software, the researchers in [17] simulated an 11-node feeder, analyzing the susceptibility of the method to different types of faults based on fault distance, resistance, and fault inception angles. The authors proposed an impedance-based alternative for locating faults in power distribution networks in the presence of distributed photovoltaic generation resources. Voltage and current measurements at the feeder and distributed generation terminals, combined with the Pi line model, were utilized to improve accuracy.

In [18], alternatives were presented for locating single-phase faults in power distribution systems with distributed generation by applying impedance-based methods. The study analyzed the influence of parameters such as the magnitude of distributed generation, fault impedance, and its relative position. In [19], reactance graphs were used for fault location in distribution systems. These graphs are defined by calculating a distance that depends on voltage and current values during fault and pre-fault conditions. This method requires the series impedance of the line in each section, as well as the fundamental component of voltage and current.

The authors in [20] implemented a fault location method based on impedance estimation, considering the effects of distributed generation and variations in load current according to voltage and current measurements from generation centers. The approach was tested using fault resistances ranging from 0 to 40 ohms and a distributed generation contribution of 5%–50%, simulating different types of single-phase, two-phase, and three-phase faults. Other methods include knowledge-based approaches, classified into three groups: artificial intelligence and statistical data analysis methods, distributed device-based methods, and hybrid methods.

In [21], an algorithm was developed in MATLAB for fault detection and classification due to short circuits, implemented in a modified IEEE 34-bus system simulated in ATPDraw. The simulations were conducted considering three scenarios, with and without distributed generation units. Additionally, the study considered factors such as fault type, incidence angle, and fault resistance. This method consists of two stages: detection and classification. In the detection stage, three-phase currents were analyzed, extracting their characteristics using the discrete wavelet transform with maximum overlap. During classification, the faulty phases were identified using three fuzzy inference systems (FIS). The results showed that the algorithm was highly effective, achieving a fault detection accuracy of over 94.9% and a classification accuracy of 100%.

In [22], the k-means algorithm was implemented in MATLAB/Simulink to detect, classify, and locate faults in the power system of the IEEE 14-bus test network. A matrix was used to record data under normal operation, in the presence of faults, and for testing. These datasets were compared, and the k-means algorithm was applied for classification. When a fault occurred, the algorithm identified the most similar row in the training data matrix, determining the type and location of the fault based on the known faults for each bus. If no fault was detected in the input row, the classifier operated normally, with the first row representing a fault-free state.

In [23], the authors used the discrete wavelet transform (DWT) combined with artificial neural networks (ANNs) to propose a method for determining the fault section and its location in power distribution systems. The tests were conducted using an IEEE 34-bus test feeder. The DWT was applied to analyze and extract the characteristics of transient signals observed

in the fault from the three-phase line current measurement at a single point. Artificial neural networks were employed to classify and predict the fault section and location, using entropy indices obtained from the decomposition of the discrete wavelet transform as input features.

Finally, in [24], the authors presented a solution to the problem of service continuity by applying knowledge-based methods using the k-means algorithm and the finite mixture technique. A mathematical model was developed based on the magnitude of voltage sags occurring during faults in a 25 kV power distribution system. By defining characteristic groups, the information was optimized, ensuring greater accuracy in the model.

This literature review highlights the significance of the problem addressed in this research regarding fault detection in power distribution systems. It demonstrates the interest of various researchers in proposing solutions based on different techniques, with clustering algorithms standing out among them.

B. Data characterization by clustering

Clustering is a data reduction process in which individuals are grouped, in contrast to principal component analysis or canonical analysis, where reduction consists of variable combinations. Therefore, clustering can be defined as the generation of a classification or typology of elements [25]. The applications of the k-means algorithm range from astronomy to bioinformatics, bibliometrics, and pattern recognition. In [26], text mining techniques and the k-means algorithm were used to generate groups of similar news article headlines. The extraction of news headlines and links from different sources was performed using an XML file.

The authors in [27] applied multicriteria decision-making methods along with clustering algorithms for financial risk analysis, using clustering validity indices on credit risk and bankruptcy information sets. Multi-objective discrete particle swarm optimization methods have also been implemented to address the network clustering problem by adopting a decomposition mechanism and introducing a population initialization algorithm. In this regard, researchers in [28] implemented clustering algorithms for high-dimensional data, employing k-d random forests and the priority search k-means tree, as well as an algorithm for matching binary features through the search of multiple hierarchical clustering trees.

The researchers in [29] developed a statistical model to estimate the proportions of genomic choices in a diverse population by applying the EM algorithm and next-generation sequencing data. Their objective was to better understand mutation distributions and the behavior of subpopulations within bacterial populations. This approach allowed the authors to assess the progression of antibiotic resistance and the alteration of resistance genes within populations.

Another application of the EM algorithm is in the field of oncology, as demonstrated in the study conducted in [30], where classical survival estimators were determined, and the relationship between COVID-19 and cancer-related deaths was analyzed. The researchers developed an algorithmic method that incorporates COVID-19-related deaths into the observed data. Additionally, the k-means and Gaussian Mixture Model-Expectation Maximization algorithms have been used to segment brain lesion areas in magnetic resonance imaging, utilizing lesion segmentation data from ischemic stroke patients [31]. These studies highlight the significance of clustering algorithms and their application across various fields of knowledge.

III. METHODOLOGY

In the first phase, the database was generated with short-circuit current values at different percentage distances from the fault point using a simulation tool, in this case, Neplan version 5.5 [32], which was used to design the distribution system. Various simulations of single-phase, two-phase, and three-phase short circuits were executed, and the information from each was extracted.

In the second phase, the number of clusters was determined using the Calinski-Harabasz index, with support from the Python programming language to obtain optimal data. This index evaluates the degree of dispersion between clusters, being directly proportional to the covariance between clusters and inversely proportional to the intra-cluster covariance, as defined in equations 1, 2, and 3.

$$\mathrm{CH}(k) = \frac{A(k)(N-k)}{D(k)(k-1)} \quad (1)$$

$$A(k) \sum_{k=1} a_k \| \, x_k - x \, \|^{2^1} \quad (2)$$

$$D(k) = \sum_{k=1} (\sum_{c(j)=k} \| \, x_j - x_k \, \|^2) \quad (3)$$

Where:
c_i: number of clusters
a: average distance of sample i to other samples in the group
A(k): covariance between clusters
D(k): intra-cluster covariance
CH: Calinski-Harabasz index
N: number of samples

The higher *A(k)*, the greater the dispersion between groups, and the smaller *D(k)*, the stronger the relationship between clusters. The higher this ratio, the greater the value of the Calinski-Harabasz index.

In the third phase, characteristic zones and homogeneity relationships of short-circuit currents were generated based on voltage drops, short-circuit currents, and selected distances in the tests using the k-means algorithm, considering the following:

Given an initial set of centroids $c_1^{(1)}$, …, $c_k^{(1)}$, the algorithm is represented in two stages:

Assignment: Each data point is assigned to the group with the closest mean, as defined in Equation 4.

$$G_i\left(^t\right) = \left\{ xp : \| \, xp - c_i(^t) \, \| \leq \| \, xp - c_j(^t) \, \| \quad \forall 1 \leq j \leq k \right\} \quad (4)$$

Where:
G_i = Group i
x_p = Set of observations
c_i = Number of centroids i.

Update: Calculation of the new centroids, as defined in Equation 5.

$$c_i\left(^{t+1}\right) = \frac{1}{|G_i(^{t+1})|} \sum_{xj \in G_i(^{t+1})} xj \quad (5)$$

In the fourth phase, the GMM algorithm was implemented based on the EM algorithm to estimate parameters such as the mean and covariance of the generated groups, with the support of the Python programming language.

In the final phase, an approximation of the probability density function of each cluster was established to determine the most probable fault zone, resulting in the probability of belonging to each one. These algorithms consider the following:

$$P(X, C|\pi, \mu, \sigma) = P(X, C, \mu, \sigma) \cdot P(C|\pi) \quad (6)$$

Where:
P = probability of an object belonging to each cluster
μ = mean of the objects in the cluster
σ = standard deviation
π = probability of belonging to a cluster

The goal is to determine the probability of belonging to a cluster, defined by Equation 6.

$$P(X, C, \mu, \sigma) \quad (7)$$

Where $P(C \mid \pi)$ is a constant value: $P(c=\mid \pi_1)$, $(P(c=2 \mid \pi_2) \ldots P(c=k \mid \pi_K)$. Thus, Equations 6 and 7 are simplified into Equation 8.

$$P(X \left| \pi, \mu, \sigma) = \sum_{k=1} K\, P(X, c = k \right| \pi, \mu, \sigma) = \sum_{k=1} K\, \pi_k \cdot P_k(X \left| \mu, \sigma) \right. \quad (8)$$

Applying the probability density function of a normal distribution, the summation of the probabilities of an object belonging to a cluster is expressed by Equations 9 and 10.

$$P\left(X \left| \pi, \mu, \sigma \right. \right) = \sum (k = 1 \text{ to } k) \pi_k \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k}} \quad (9)$$

$$P\left(X \mid \pi, \mu, \sigma \right) = \sum (k = 1 \ldots k) \frac{\pi_k e^{-\frac{(x-\mu_k)^2}{2\sigma_k}}}{\sigma_k \sqrt{2\pi}} \quad (10)$$

Representing Equation 11, the highest probability of belonging to a cluster.

$$\operatorname{argmax} \sum (k = 1 \ldots k) \frac{\pi_k e^{-\frac{(x-\mu_k)^2}{2\sigma_k}}}{\sigma_k \sqrt{2\pi}} \quad (11)$$

Knowing the parameters of the distributions and the probability π of belonging to a cluster, each object is assigned to the one with the highest probability of membership.

$$\operatorname{argmax} \sum (k = 1 \ldots k) \pi_k P_k(X, \mu, \sigma) \quad (12)$$

Finally, the parameters μ, π, and σ are calculated according to equations 13, 14, and 15.

$$\pi_K = \frac{\text{Number of objects in the cluster}}{\text{Total number of objects}} \quad (13)$$

$$\mu_k = \frac{1}{n} \sum (k = 1 \ldots k) x_i \quad (14)$$

$$\sigma_k = \sqrt{\frac{1}{n-1} \sum (n = 1 \ldots n) (x_i - \mu_k)^2} \quad (15)$$

IV. RESULTS

In Fig 1, the medium-voltage distribution system operating at 34.5 kV and 0.44 kV, which was implemented in this research, is shown. This system consists of a source or network equivalent (F-17) with a capacity of 477.03 MVA and two three-winding transformers rated at 34.5/0.44 kV. From the secondary windings of these transformers, two lines extend from each, with a total of 15 connected loads.

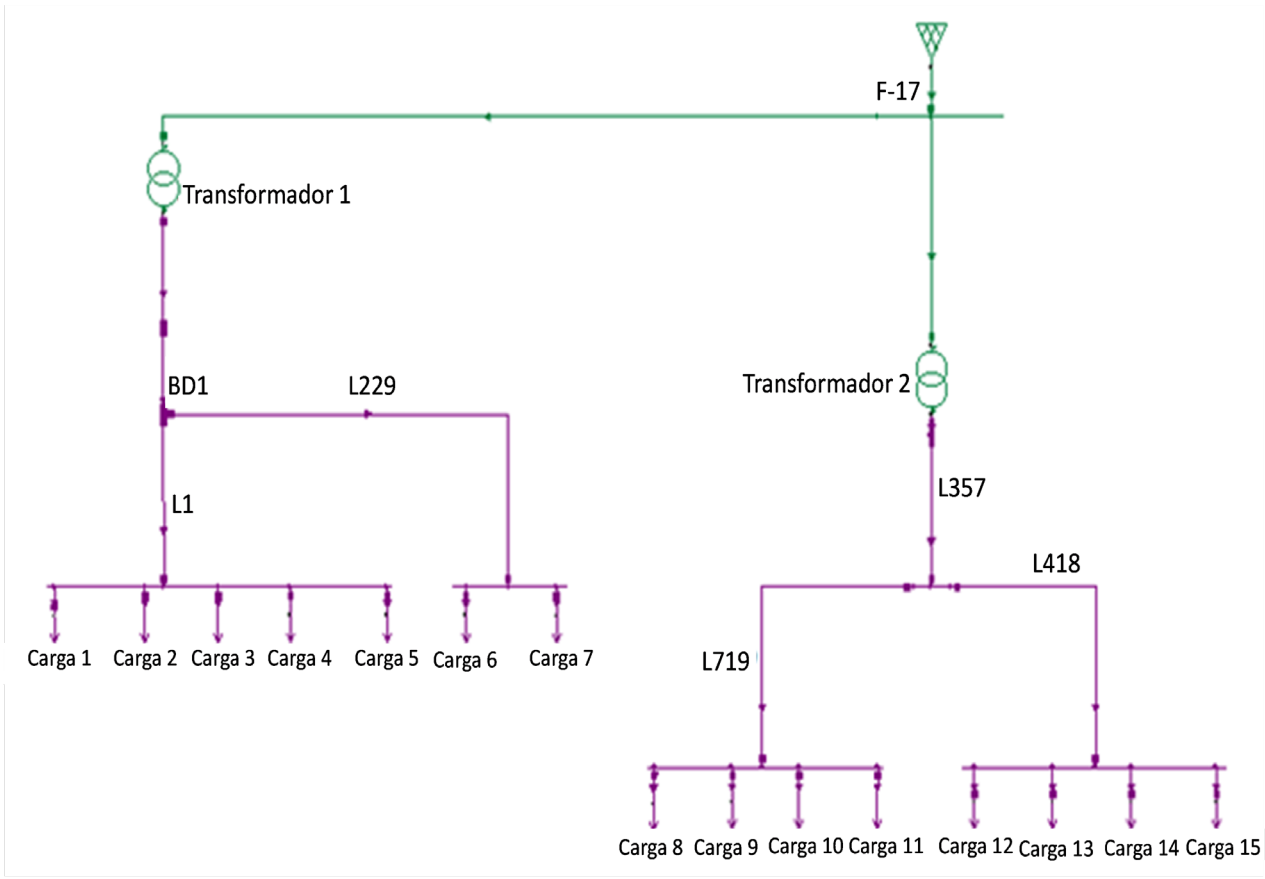

Fig 1. Single-line diagram of the 34.5 kV power distribution system. Source: Authors

Table 1 presents the parameters of line L719, including apparent, active, and reactive power, as well as current and power factor.

TABLE 1. ELECTRICAL PARAMETERS OF THE LOADS ON LINE L719

Load	Apparent Power (KVA)	Active Power (KW)	Reactive Power (kVAr)	Current (kA)	Power Factor
8	0.049	0.037	0.032	0.062	0.75
9	0.111	0.083	0.073	0.141	0.75
10	0.011	0.009	0.008	0.015	0.75
11	0.122	0.091	0.081	0.156	0.75

Source: Authors

Table 2 records the parameters of line L418, including apparent, active, and reactive power, as well as current and power factor.

TABLE 2. ELECTRICAL PARAMETERS OF THE LOADS ON LINE L418

Load	Apparent Power (KVA)	Active Power (KW)	Reactive Power (kVAr)	Current (kA)	Power Factor
12	0.022	0.017	0.015	0.028	0.75
13	0.076	0.057	0.050	0.097	0.75
14	0.046	0.034	0.030	0.058	0.75
15	0.046	0.034	0.030	0.058	0.75

Source: Authors

The faults were executed on lines L1, L229, L357, L419, and L719, which were subjected to simulations to obtain the short-circuit fault current at different distance percentages. The distances ranged from 2% to 98% since the software does not allow simulations from 0% to 100%. The short-circuit mode was used to simulate single-phase, two-phase, and three-phase faults, selecting the fault type as shown in Fig 2. It is worth mentioning that the calculation method used was IEC60909:2001.

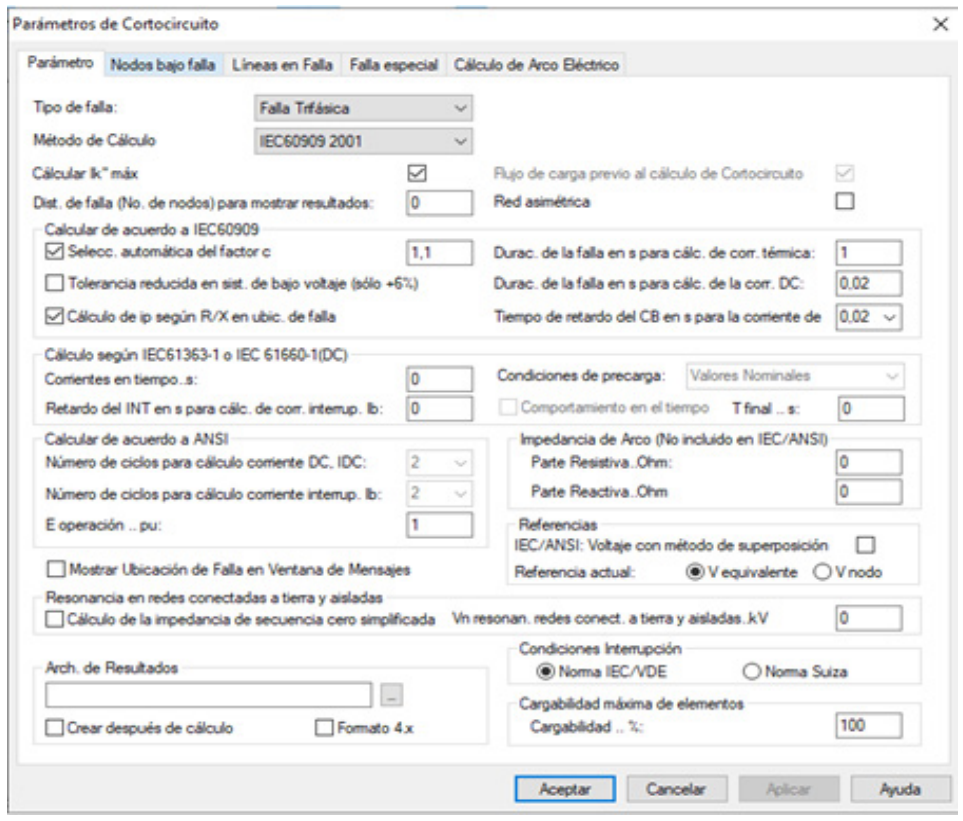

Fig 2. Fault type selection. Source: Authors

When extracting the results from each simulation, the nominal voltage, line-to-ground fault voltage, line-to-ground fault voltage angle, short-circuit current (Ik"), and short-circuit current angle for each line were obtained. However, only the percentage distance and short-circuit current were used in this study, as the other variables did not exhibit variations significant enough to describe the behavior of the lines under fault conditions for clustering modeling. Once this database was extracted, it was subjected to the generation of random values based on its characteristics for subsequent processing and clustering formation.

Table 3 presents the single-phase fault short-circuit currents obtained from the simulations. The complete database and Python codes are available in the GitLab repository at the following link: https://gitlab.com/daniela-castillo/extraccion-de-bd-de-fallas-por-cortocircuito.

TABLE 3. SAMPLE OF SINGLE-PHASE FAULT SHORT-CIRCUIT CURRENTS

Distance (%)	Ik" (kA)
2	50.9
2	33.5
4	50.5
4	32.4
4	32.0
6	50.1
6	31.4

Source: Authors

Table 4 presents some of the short-circuit current values for two-phase faults extracted from the simulations. The complete dataset is available in the previously mentioned repository, along with the results of the three-phase fault simulations.

TABLE 4. SHORT-CIRCUIT CURRENTS FOR TWO-PHASE FAULTS

Distance (%)	Ik" (kA)
2	43.0
2	43.0
2	32.8
4	42.7
4	42.7
4	32.1
6	42.5
6	42.5
6	31.5

Source: Authors

These data were uploaded to a Google Colab notebook [33], and the reading process was performed using Python. After extracting the data, the optimal number of clusters was determined using the Calinski-Harabasz criterion to optimize the classification algorithm, as the k-means algorithm requires this value to be specified, either through visual inspection or based on the analyst's decision. Fig 3 shows the result of applying this criterion to the data extracted from single-phase faults, indicating that two clusters should be selected. The same result was obtained for two-phase and three-phase faults.

Once the number of clusters for each fault type was determined, the centroids were initialized at random points. The distance of each point to the centroid was calculated, and the points were grouped with the nearest centroid. Subsequently, the centroids were recalculated iteratively until they stabilized as the centers of each cluster. Fig 4 shows the result of applying the k-means algorithm, where two fault zones are defined. These zones exhibit low data dispersion until reaching a certain distance.

For all three types of faults, the formation of two potential fault zones was observed. Upon reviewing the single-line diagram, it was identified that lines L1, L229, L719, and L418 are located below a node positioned after the transformers. According to Kirchhoff's circuit law, the current passing through this node splits into two paths, reducing its magnitude compared to the current flowing through line L357. In the same figure, point (2;34) is plotted, representing data from a fault occurring in line L357. When applying the clustering algorithm, the highest probability of belonging is assigned to zone one, validating the accuracy of the algorithm.

```
archivo= 'Monofásicas.xlsx'
df=leerExcel(archivo)

optimal_clusters= calinski_harabasz(data
print ("Número optimo de clusters:", opt

/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/
  warnings.warn(
Número optimo de clusters: 2
```

Fig 3. Calinski-Harabasz index applied to single-phase short-circuit faults. Source: Authors

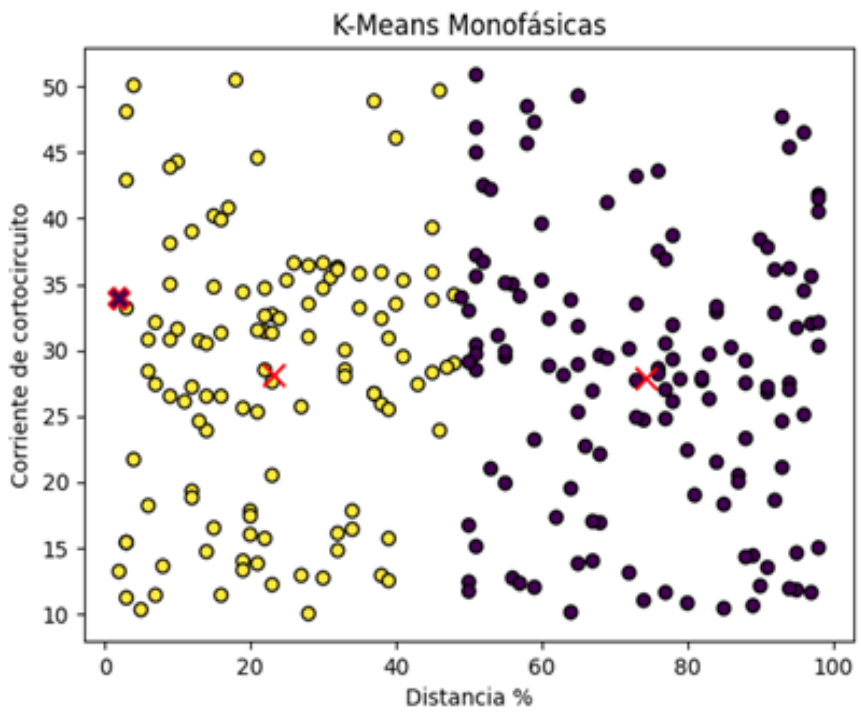

Fig 4. Application of the k-means algorithm to the extracted data from single-phase faults. Source: Authors

Once the k-means algorithm was implemented, the GMM algorithm based on the EM algorithm was implemented to obtain the covariance and mean values for each centroid. These are key parameters in mixture models and the estimation of the probability density function for cluster membership. As a final result, the probability of belonging to each cluster was generated.

Table 5 presents the results of applying these algorithms, where the means or centroids are located at coordinates (18.600; 28.919) and (68.149; 25.109). This means that data points closer to one of these centroids will belong to that cluster. Based on the covariance values, the variability range for the first cluster in terms of distance percentages (represented by x) extends from 159.29 to 33.66, while for the second cluster, it ranges from 33.66 to 84.08.

Regarding short-circuit currents, the range for the first cluster is from 285.07 to -2.96, while for the second, it extends from -2.96 to 87.21. In terms of probabilities, the second cluster exhibits a higher probability, indicating that, based on the extracted data, faults are more likely to occur in zone two. However, the probability difference between both zones is minimal, with only a 0.994847% difference. This is expected since the current values used in the simulation do not vary significantly from one zone to another due to the system's voltage level.

TABLE 5. RESULTS OF THE GMM-EM ALGORITHM IMPLEMENTATION

	Cluster 1		Cluster 2	
Centroids	(18.60035532; 28.91901509)		(68.1491318; 25.10959859)	
Probabilities	0.45025765		0.54974235	
Covariance				
Distance (%)	159.2908293	33.66178475	33.66178475	84.08520891
Ik" (kA)	285.0738741	-2.96782379	-2.96782379	87.2189548

Source: Authors

In Fig 5, the graphical representation of the application of the GMM algorithm based on EM is shown. It can be observed that the largest volume of data is in zone two, represented by the yellow color, where both zones are delimited by distance ranges. In a test conducted by locating a point based on a fault occurring on line L357, it was correctly placed in zone one.

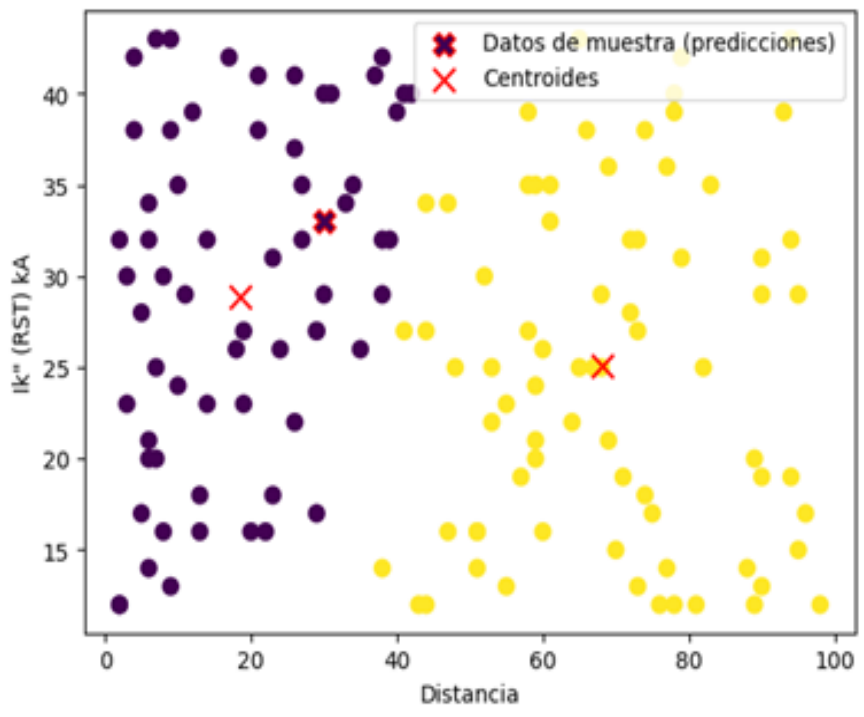

Fig 5. Graphical representation of the application of the GMM-EM algorithm. Source: Authors

A validation of the algorithms was performed using the data extracted from the initial simulations, and different points were located using the same dataset from the development. A sample of this validation is presented in Table 6 for single-phase faults occurring in zone one, with four test cases.

TABLE 6. SAMPLE OF SHORT-CIRCUIT CURRENTS IN SINGLE-PHASE FAULTS

Coordinates	Results
(30; 45.367)	Located in Zone 1
(44; 42.852)	Located in Zone 1
(54; 41.168)	Dispersion identified, assigned to Zone 2
(88; 36.117)	Located in Zone 2

Source: Authors

In Fig 6, the results of the tests performed on single-phase short-circuit faults are shown. In Fig 6(a), the fault is correctly located at point (30; 45.367), meaning it is closer to the centroid of Zone 1. Similarly, in Fig 6(b), the fault is also correctly positioned at point (44; 42.852), which is closer to the centroid of Zone 1. Finally, in cases where the fault approaches the distance ranges belonging to the other zone, these faults are incorrectly assigned, as observed in Fig 6(c) and Fig 6(d) at points (54; 41.168) and (88; 36.117). This misclassification is due to the voltage levels used and the limitations in implementing greater variations in current between zones.

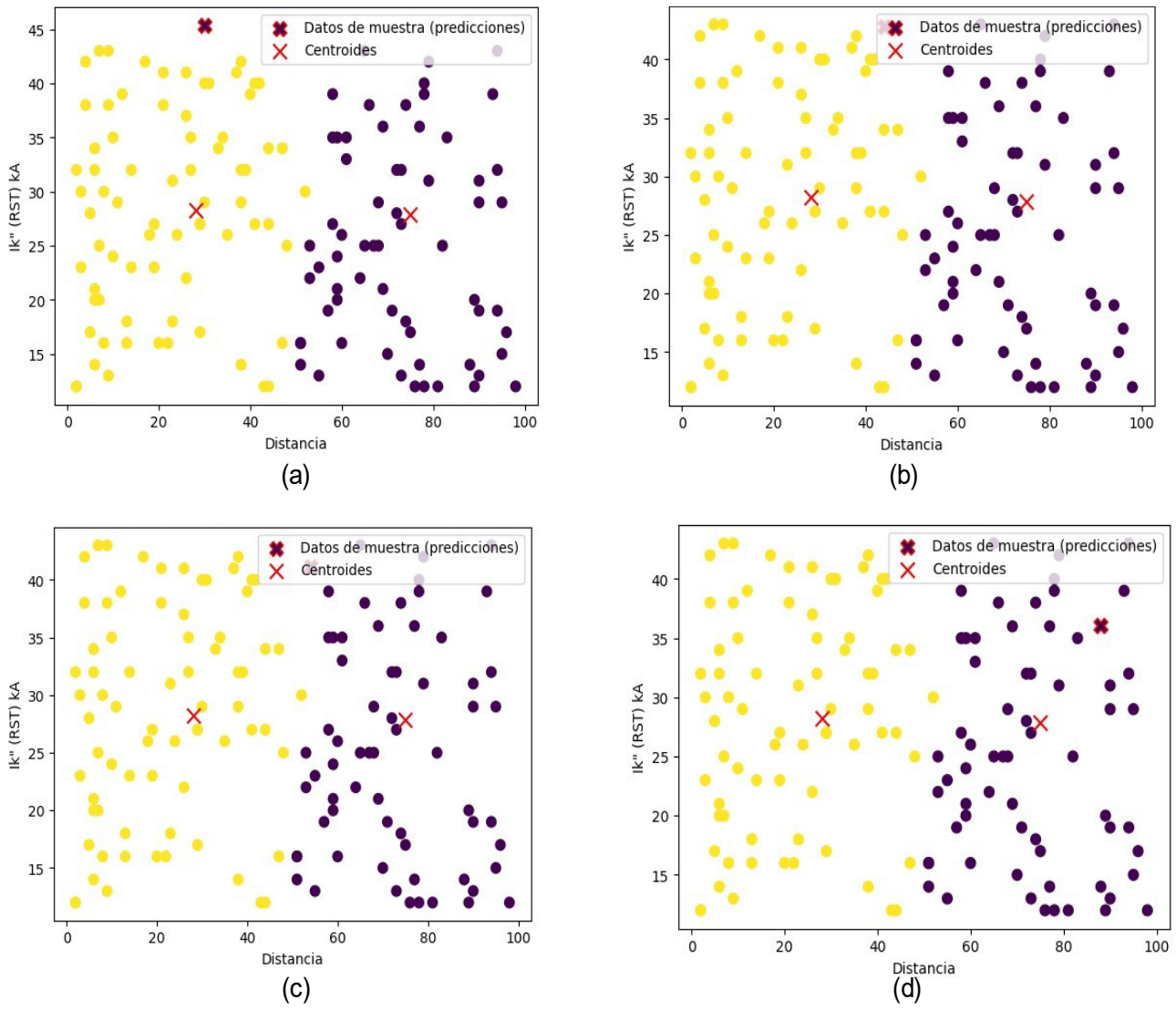

Fig 6. Results for single-phase short-circuit faults. Source: Authors

In [34], a methodology was developed to determine fault zones using a statistical model based on voltage sag databases. The authors used a database containing different short-circuit fault cases by varying the fault resistance between 0 and 50Ω. They grouped the maximum sag values in the phases of a three-phase system using the k-means algorithm. The classification process involved several steps: first, identifying the faulty phase; second, determining the fault resistance value, which allowed for the creation of scenarios grouping resistance values into specific ranges; and third, locating the fault by applying a mixture model and comparing them.

In [35], the system was divided into zones, first by branch and then into smaller sub-zones. Data acquisition incorporated both real and simulated fault histories, extracting descriptors from the RMS values of phase voltages and currents, with labels assigned according to the corresponding zone. Support vector machines were applied, initially analyzing behavior with four zones and gradually increasing to ten. The process continued with further subdivisions until results no longer aligned with previous iterations. Unlike the approach in this work, no comparison was made between different zone configurations; instead, the number of zones was determined using the Calinski-Harabasz index. Additionally, fault location relied on

support vector machines, as no probability density assumptions were required. In contrast, the present study determines fault location based on the probability of belonging to each zone.

In [36], a methodology was developed using variations in the root mean square values of current, voltage, apparent power, and reactance before and during the fault as input data, considering fault resistances ranging from 0.5 to 40Ω. The Chu-Beasley genetic algorithm was applied to select the optimal parameterization during the configuration of the K-NN-based classifier. A class was assigned to each training data point according to the zone, which was generated using the Matlab and ATP simulation tools. The authors implemented the classification algorithm to identify the fault zone, and once determined, they selected an equivalent branch to estimate the distance to the fault. With this information, the fault location was obtained by analyzing the responses of example-based learning methods and fault reactance estimation.

Later, in [37], a methodology was developed for fault location in distribution systems, but instead of using distances to the fault zone, it considered variations in loads, substation voltages, and line lengths. However, it shared the approach of defining zones and node distances for database construction. The same clustering algorithm used in the previous work was implemented to estimate the fault occurrence distance. Additionally, the "Boosting" classification method was implemented to determine the fault zone. Unlike the methodology presented in this work, zone classification was performed using an unsupervised learning algorithm such as k-means, along with a Gaussian mixture model based on the expectation-maximization algorithm to establish the probability of zone membership.

In [38], a methodology was proposed for fault location in a distribution system based on evaluating the phases followed by the maintenance operator handling the event and proposing solutions. During the identification or segmentation process, current and voltage values were used to determine fault states by comparing them with events exhibiting specific waveform patterns. A metaheuristic optimization method was applied to compare calculated values with those obtained from industrial distribution system simulations. Finally, the causes of faults were characterized to classify them.

Unlike the previously mentioned methodologies, in this work, the number of zones is determined using the Calinski-Harabasz index, ensuring an optimal number of fault zones. No comparisons are made by varying the number of zones or directly modifying fault resistances. However, like other works, this study involves extracting variables that enable zone labeling based on their characteristics. In some cases, supervised learning classification algorithms were used, but zone estimations were performed through data analysis.

V. CONCLUSIONS

With the application of the k-means and GMM algorithms, it is possible to identify potential fault zones based on the data extracted from the tests. Although a single fault zone is not explicitly determined, since the algorithms perform estimations based on the processed data, these tools still support decision-making based on the obtained estimations. The variation in the distance percentages to the fault line helps identify possible fault zones. However, greater accuracy could be achieved if the software allowed for the variation of additional parameters to model a larger number of clusters with more precise characteristics.

The implementation of the Calinski-Harabasz algorithm allowed for determining a different number of groups than initially estimated through visual inspection. However, after its implementation, it became clear why this number was chosen, considering that only short-circuit distance and current values were used, with no significant variations between zones. The proposed methodology achieved its main objective: detecting potential fault-affected zones in a medium-voltage power distribution system. By applying machine learning techniques, it was possible to identify fault zones and infer that this approach could be extended to higher-voltage distribution systems. However, the grouping and behavior of the lines in such systems could result in a greater number of fault zones.

As future work, it is proposed to simulate a higher voltage substation (115 kV), which is expected to provide a wider range of short-circuit fault currents for testing and comparison. The applied methodology successfully identified the zone with the highest probability of

failure based on current and distance percentage. However, further studies could focus on predictive algorithms applicable to this case, enabling fault zone prediction using only current measurements.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

D. Castillo-Acosta: Investigation, Software, Data Curation, Writing – Original Draft. **C. Robles-Algarín:** Methodology, Validation, Writing – Review and Editing, Supervision. **L. Camargo-Ariza:** Validation, Visualization, Supervision.

REFERENCES

[1] Comisión de Regulación de Energía y Gas, "Identificación y análisis de los niveles de calidad del servicio alcanzables en las redes de distribución de energía eléctrica del SIN," Apolo CREG, Dec. 2013. http://apolo.creg.gov.co/Publicac.nsf/52188526a7290f8505256eee0072eba7/c55a6288b2a5fd1a05257cfb0054ae53/$FILE/Circular036-2014%20Anexo.pdf (accessed Jul. 31, 2023).

[2] Comisión de Regulación de Energía y Gas, "Resolución No. 070 de 1998," Apolo CREG, Oct. 2008. http://apolo.creg.gov.co/Publicac.nsf/Indice01/Resoluci%C3%B3n-1998-CREG070-98 (accessed Jul. 31, 2023).

[3] Superintendencia delegada para Energía y Gas, "Diagnóstico de la calidad del servicio de energía eléctrica en Colombia 2017," Superservicios.gov, Jul. 2018. https://www.superservicios.gov.co/sites/default/files/inline-files/diagnosticodecalidad2017_version_06072018_1%20%281%29.pdf (accessed Jul. 31, 2023).

[4] Superintendencia delegada para Energía y Gas, "Diagnóstico de la calidad del servicio de energía eléctrica en Colombia 2018," 2019.

[5] D. Vargas Jiménez, "Estudio de indicadores de calidad del suministro de energía eléctrica de una red de distribución a 13.2 kV," 2015. Accessed: Jul. 31, 2023. [Online]. Available: https://repositorio.unal.edu.co/handle/unal/60037

[6] J. Dolores and J. Cervantes, Sistemas de distribución de energía eléctrica / José Dolores Juárez Cervantes. 2002. https://core.ac.uk/download/pdf/48392416.pdf

[7] C. J. Zapata G, "Colapsos de voltaje en sistemas de potencia," ACIEM, May 1996. https://academia.utp.edu.co/planeamiento/files/2014/01/cjz-colapsos-voltaje.pdf

[8] J. J. Mora Flórez, J. Colomer L. Colomer, Joaquim. Meléndez Frigola, Gilberto. Carrillo Caicedo, and I. i A. Universitat de Girona. Departament d'Electrònica, Localización de faltas en sistemas de distribución de energía eléctrica usando métodos basados en el modelo y métodos basados en el conocimiento. Universitat de Girona, 2007.

[9] A. O'Hagan, T. B. Murphy, L. Scrucca, and I. C. Gormley, "Investigation of Parameter Uncertainty in Clustering Using a Gaussian Mixture Model Via Jackknife, Bootstrap and Weighted Likelihood Bootstrap," Oct. 2015, [Online]. Available: http://arxiv.org/abs/1510.00551

[10] Y. Xie, S. Zhou, Y. Xiao, S. Kulturel-Konak, and A. Konak, "A ß-accurate linearization method of Euclidean distance for the facility layout problem with heterogeneous distance metrics," *Eur J Oper Res*, vol. 265, no. 1, pp. 26–38, Feb. 2018, doi: 10.1016/J.EJOR.2017.07.052.

[11] A. García Herrero, "Algorithms for Gaussian mixture model(gmm) estimation," 2015. Accessed: Jul. 31, 2023. [Online]. Available: https://repositorio.unican.es/xmlui/bitstream/handle/10902/6676/376307.pdf?sequence=1

[12] M. Gong, Q. Cai, X. Chen, and L. Ma, "Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 1, pp. 82–97, Feb. 2014, doi: 10.1109/TEVC.2013.2260862.

[13] M. Charrad, N. Ghazzali, U. Laval, and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set Véronique Boiteau," 2014. [Online]. Available: http://www.jstatsoft.org/

[14] B. Desgraupes, "Clustering Indices," 2017.

[15] M. Mirzaei, Z. Kadir, and H. Hizam, "Lightning Measurement involving 4 base stations located at UTM Skudai, UTeM Melaka, Kolej Uniti Negeri Sembilan and Universiti Tenaga Nasional (UNITEN) View project Heuristic Incipient Fault Monitoring and Diagnostic Platform for Line Start Permanent Magnet Synchronous Motors View project," 2009. [Online]. Available: https://www.researchgate.net/publication/234682955

[16] O. O. Austin, O. Alonge, and A. J. Adeniyi, "Fault Diagnosis Algorithm and Protection of Electric Power Systems in an Alternative Distribution System," *Journal La Multiapp*, vol. 1, no. 3, pp. 8–16, Dec. 2020, doi: 10.37899/journallamultiapp.v1i3.192.

[17] R. Dashti, M. Ghasemi, and M. Daisy, "Fault location in power distribution network with presence of distributed generation resources using impedance based method and applying π line model," *Energy*, vol. 159, pp. 344–360, Sep. 2018, doi: 10.1016/J.ENERGY.2018.06.111.

[18] J. Faig, J. Melendez, S. Herraiz, and J. Sánchez, "Analysis of faults in power distribution systems with distributed generation," *Renewable Energy and Power Quality Journal*, vol. 1, no. 8, pp. 863–868, Apr. 2010, doi: 10.24084/repqj08.502.

[19] G. Morales-España, H. Vargas, and J. Mora-Florez, "Método de localización de fallas en sistemas de distribución basado en gráficas de reactancia," Scientia Et Technica, Aug. 2007.

[20] W. J. Gil-González, J. J. Mora-Flórez, and S. Pérez-Londoño, "Comparative analysis of metaheuristics optimization techniques to parameterize fault locators for power distribution systems," 2013. Accessed: Jul. 31, 2023. [Online]. Available: http://www.scielo.org.co/pdf/inco/v15n1/v15n1a10.pdf

[21] A. da S. Santos, L. T. Faria, M. L. M. Lopes, A. D. P. Lotufo, and C. R. Minussi, "Efficient Methodology for Detection and Classification of Short-Circuit Faults in Distribution Systems with Distributed Generation," *Sensors*, vol. 22, no. 23, Dec. 2022, doi: 10.3390/s22239418.

[22] A. Recioui, B. Benseghier, and H. Khalfallah, "Power system fault detection, classification and location using the K-Nearest Neighbors," Aug. 2015, pp. 1–6. doi: 10.1109/INTEE.2015.7416832.

[23] A. C. Adewole, R. Tzoneva, and S. Behardien, "Distribution network fault section identification and fault location using wavelet entropy and neural networks," *Appl Soft Comput,* vol. 46, pp. 296–306, Sep. 2016, doi: 10.1016/J.ASOC.2016.05.013.

[24] J. Mora-Flórez, J. Cormane-Angarita, and G. Ordóñez-Plata, "k-means algorithm and mixture distributions for locating faults in power systems," *Electric Power Systems Research*, vol. 79, no. 5, pp. 714–721, May 2009, doi: 10.1016/j.epsr.2008.10.011.

[25] I. Olkin and A. R. Sampson, "Multivariate Analysis: Overview," *International Encyclopedia of the Social & Behavioral Sciences*, pp. 10240–10247, Jan. 2001, doi: 10.1016/B0-08-043076-7/00472-1.

[26] P. Lama, "Bachelor's thesis (UAS) Information Technology Text Mining and Clustering 2013 Prabin Lama CLUSTERING SYSTEM BASED ON TEXT MINING USING THE K-MEANS ALGORITHM 45 pages Instructor: Patric Granholm CLUSTERING SYSTEM BASED ON TEXT MINING USING THE K-MEANS ALGORITHM," 2013.

[27] G. Kou, Y. Peng, and G. Wang, "Evaluation of clustering algorithms for financial risk analysis using MCDM methods," *Inf Sci (N Y)*, vol. 275, pp. 1–12, Aug. 2014, doi: 10.1016/j.ins.2014.02.137.

[28] M. Gong, Q. Cai, X. Chen, and L. Ma, "Complex Network Clustering by Multiobjective Discrete Particle Swarm Optimization Based on Decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 1, pp. 82–97, 2014, doi: 10.1109/TEVC.2013.2260862.

[29] Y. Zhang et al., "An expectation–maximization algorithm for estimating proportions of deletions among bacterial populations with application to study antibiotic resistance

gene transfer in Enterococcus faecalis," *Mar Life Sci Technol*, vol. 5, no. 1, pp. 28–43, Feb. 2023, doi: 10.1007/s42995-022-00144-z.

[30] F. De Felice, L. Mazzoni, and F. Moriconi, "An Expectation-Maximization Algorithm for Including Oncological COVID-19 Deaths in Survival Analysis," *Current Oncology*, vol. 30, no. 2, pp. 2105–2126, Feb. 2023, doi: 10.3390/curroncol30020163.

[31] J. Qiao et al., "Data on MRI brain lesion segmentation using K-means and Gaussian Mixture Model-Expectation Maximization," *Data in Brief*, vol. 27, Dec. 2019, doi: 10.1016/j.dib.2019.104628.

[32] PSI Neplan AG. NEPLAN. Archivo (Versión 5.5) 2012. Accessed: Sep. 05 https://www.neplan.ch/archive/?lang=es

[33] Google Research. Google Colab. Ap. 2018 Accessed: Sep. 05 https://colab.research.google.com/?hl=es

[34] Ordóñez Plata, G., Angarita, J. C., & Mora Flórez, J. "Faulted zone determination using statistical modeling of voltage sag database in power distribution systems", *Rev. Fac. Ing. Univ. Antioquia*, no. 47, pp. 197-208, 2009.

[35] Morales España, G., Barrera Cárdenas, R., Raúl, H., & Torres, V. "Unique localization of faults in distribution systems by means of zones with SVM", *Rev. Fac. Ing. Univ. Antioquia*, no. 47, pp. 187-96, 2009.

[36] Jagua Gualdrón, J. L. Metodología para la localización de fallas en sistemas de distribución. 2022. Available: https://repositorio.unal.edu.co/bitstream/handle/unal/81834/1098607872.2022.pdf?sequence=3

[37] Zapata Tapasco, A., Pérez Londoño, S., & Mora Flórez, J. "Metodología híbrida basada en el regresor knn y el clasificador boosting para localizar fallas en sistemas de distribución", *Ingeniería y competitividad*, vol. 16, no. 2, pp. 165–177, 2014.

[38] Zapata Tapasco, A., Pérez Londoño, S., & Mora Flórez, J. "Método basado en clasificadores k-NN parametrizados con algoritmos genéticos y la estimación de la reactancia para localización de fallas en sistemas de distribución", *Rev. Fac. Ing. Univ. Antioquia*, no. 70, pp. 220-232, 2014.

Daniela Castillo-Acosta received her Electronic Engineering degree from the Universidad del Magdalena, Santa Marta, Colombia, in 2019. She later obtained a Master's degree in Engineering from the same university in 2024. Her professional experience has been focused on the energy sector, starting in the commercial area and later working as a professional in Advanced Metering Infrastructure (AMI). She is currently a lead professional in the billing area at Compañía Energética de Occidente, Popayán, Cauca. https://orcid.org/0009-0007-5745-159X

Carlos Robles-Algarín received his Master's degree in Control Engineering and his Ph.D. in Technology Management from Universidad Dr. Rafael Belloso Chacín, Maracaibo, Venezuela, in 2011 and 2017, respectively. In 2004, he obtained his Electronic Engineering degree from Universidad del Norte, Barranquilla, Colombia. He is currently a Full Professor at Universidad del Magdalena, where he teaches in the Electronic Engineering, Master's, and Doctoral programs in Engineering. He is the leader of the Magma Ingeniería Research Group, categorized as A1, and is classified as a Senior Researcher. https://orcid.org/0000-0002-5879-5243

Luis Camargo Ariza received his Doctor of Science degree in 2019 from Universidad Dr. Rafael Belloso Chacín, Maracaibo, Venezuela. In 2009, he obtained his Master's degree in Electronic Engineering from Universidad Nacional Experimental del Táchira, San Cristóbal, Venezuela. In 2003, he earned his Electronic Engineering degree from Universidad Francisco de Paula Santander, Cúcuta, Colombia. He is a Full Professor at Universidad del Magdalena, a Senior Researcher, and a member of the Research Group on Electronic Development and Mobile Applications (GIDEAM), category A. https://orcid.org/0000-0002-7956-441X