LOCALIZACIÓN DE FALLAS: HERRAMIENTA DE CLASIFICACIÓN BASADA EN MEZCLAS FINITAS

LUIS ENRIQUE LÓPEZ RUIZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES
BUCARAMANGA
2007

LOCALIZACIÓN DE FALLAS: HERRAMIENTA DE CLASIFICACIÓN BASADA EN MEZCLAS FINITAS

Proyecto de Grado en modalidad de investigación presentado como requisito para optar al título de Ingeniero Electricista

LUIS ENRIQUE LÓPEZ RUIZ

Director: Prof. Dr. HERMANN RAUL VARGAS TORRES
Codirector: Ing. JORGE ANDRÉS CORMANE ANGARITA

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES
BUCARAMANGA
2007

Resumen

TÍTULO[*]
LOCALIZACIÓN DE FALLAS: HERRAMIENTA DE CLASIFICACIÓN BASADA EN MEZCLAS FINITAS

AUTOR[**]
LUIS ENRIQUE LÓPEZ RUIZ

PALABRAS CLAVES
Localización de fallas, Modelo estadístico, Clasificación, Mezclas finitas.

DESCRIPCIÓN

Un sistema de distribución se encarga de suministrar energía eléctrica a los usuarios para su posterior aprovechamiento. En la actualidad, la electricidad se ha convertido en un artículo de primera necesidad para el común de las personas. La calidad del servicio de la energía eléctrica se traduce en la capacidad de brindar electricidad sin interrupciones a niveles de tensión y frecuencia constantes. Las interrupciones del servicio pueden aparecer a causa de daños en los equipos y redes que conforman el sistema de distribución.

Se estima que el 80 % de las interrupciones ocurren debido a fallas en las redes de distribución. Las redes de distribución de un sistema eléctrico de potencia, debido a su naturaleza, poseen una compleja estructura que dificulta su rápida revisión. Determinar el origen de una falla de forma rápida, permitiría solventar el problema al reducir el tiempo de atención de las mismas.

Este trabajo de grado presenta una solución alterna a los problemas de localización de fallas en sistemas de distribución. Se aprovecha la técnica de mezclas finitas como herramienta estadística implementada en "software", para elaborar modelos que describen el comportamiento más probable del sistema bajo condiciones de falla. El propósito es clasificar adecuadamente cada evento de falla y estimar su localización dentro de la red con base a la información contenida en las señales registradas por los equipos de medida. La utilización de esta herramienta tiene como objetivo, brindar una alternativa económica y de fácil operación, encaminada a mejorar los tiempos de atención y recuperación del sistema. Lo anterior conduce a mejorar la confiabilidad y los procedimientos de planeación y operación de la red misma.

---

[*]Trabajo de investigación o tesis.

[**]Facultad de Ingenierías Físico Mecánicas, Escuela de Ingeniería Eléctrica, Electrónica y Telecomunicaciones, Hermann Raúl Vargas Torres,Jorge Andrés Cormane Angarita.

Summary

TITTLE[*]
LINE FAULTS LOCALIZATION: A CLASIFICATION TOOL BASED ON FINITES MIXTURES.

AUTOR[**]
LUIS ENRIQUE LÓPEZ RUIZ

KEYWORDS
Line faults localization, Estadistics models, Clasification, Finites mixtures.

DESCRIPCIÓN

A distribution system takes charge to supply electric power to customers for its later use. At the present time, electricity has become an article of first necessity for the common of people. Quality of service within electric power supply is translated in the capacity to offer electricity without interruptions at constant level of voltage and frequency. Interruptions of service can appear due to damages in electric equipments and networks that make part to the distribution system.

It is estimated that 80 % of the electric interruptions happens due to line faults in the distribution systems. The nets of distribution of an electric system power, due to their nature, possess a complex structure that it hinders their quick revision. To determine the origin of a line fault in a quick way, it would allow to solve the problem, reducing the time of attention in the system.

This grade thesis presents an alternating solution to the problems of localization of line faults in distribution systems. The propose is take advantage of the finite mixtures technique as statistical tool implemented in "software", to elaborate models that describe the most probable behavior in the system on line faults conditions. The purpose is to classify each line fault event appropriately and to estimate its localization inside the distribution power system with base to the information contained in the signs registered by the measure electric devices. The use of this tool has as objective, to offer an economic alternative and of easy operation, guided to improve the times of attention and recovery of the systems. The abovementioned leads to improve the dependability and the planner procedures and operation on the distribution system.

---

[*]Trabajo de investigación o tesis.

[**]Facultad de Ingenierías Físico Mecánicas, Escuela de Ingeniería Eléctrica, Electrónica y Telecomunicaciones, Hermann Raúl Vargas Torres, Jorge Andrés Cormane Angarita.

Agradecimientos

Quiero agradecer a las personas que participaron directamente en la consecución de este trabajo, el Profesor Hermann Raúl Vargas Torres y el Ingeniero Jorge Andrés Cormane Angarita, por su orientación y asistencia.

A los Ingenieros José Antonio Álvarez Duque, Felipe Cárdenas Plata y Julio Alonso Reyes Cordero, por compartir los conocimientos adquiridos en sus carreras como profesionales.

A mis amigos y compañeros, por compartir sus ideas y brindar su apoyo incondicional.

A mi madre Zully
A mis tías Gilma y Alicia
A mi novia Edna

"*El hombre encuentra a Dios,*
detrás de cada puerta que la ciencia logra abrir"
Albert Einstein

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Un sistema de distribución se encarga de suministrar energía eléctrica a los usuarios para su posterior aprovechamiento. En la actualidad, la electricidad se ha convertido en un artículo de primera necesidad para el común de las personas. Es así, que suministrar de una manera práctica y económica el servicio de energía eléctrica a cada usuario conectado a la red de distribución, se ha convertido en un compromiso de todas las empresas que prestan dicho servicio.

La calidad del servicio de la energía eléctrica se traduce en la capacidad de brindar electricidad sin interrupciones a niveles de tensión y frecuencia constantes. Las interrupciones del servicio pueden aparecer a causa de daños en los equipos y redes que conforman el sistema de distribución.

El principal inconveniente de las empresas distribuidoras del servicio es solventar las interrupciones producidas en períodos de tiempo muy cortos.

El proyecto se realizó en varias etapas de la siguiente forma:

- **Obtención de datos:** Selección y simulación de fallas en sistema de distribución utilizando los programas ATP/EMTP y MATLAB. Descripción de los tipos de fallas, topología del circuito, modelos utilizados y parámetros de simulación.
- **Reconocimiento de patrones:** Selección de los descriptores utilizados en las mezclas.
- **Construcción del modelo de localización de fallas:** Definición de las variables de entrada y salida del modelo; parámetros del modelo para la identificación, clasificación y localización de las fallas.
- **Complementación, prueba y resultados:** Obtener información complementaria que ayude a mejorar el uso de la técnica. Aplicación de la técnica mezclas finitas.

- **Diseño del software de la herramienta:** Elaboración del software, descripción del programa y elaboración del manual del usuario.

El Grupo de Investigación en Sistemas de Energía Eléctrica *GISEL* se interesa por el desarrollo de trabajos en localización de fallas en sistemas de distribución de energía eléctrica. Este trabajo hace parte de una propuesta macro, orientada bajo la modalidad de tesis de maestría como alternativa en las investigaciones sobre métodos híbridos para localización fallas en sistemas de distribución.

Capítulo 2

Objetivos

Objetivos generales

- Aplicar la técnica estadística de las mezclas finitas a la localización de fallas en sistemas de distribución.

Objetivos específicos

- Identificar el tipo de falla a partir del análisis de las señales de tensión y corriente.
- Obtener señales de tensión y corriente a partir de la simulación de un circuito prototipo en condiciones de falla utilizando ATP/EMTP y MATLAB.
- Determinar la localización de la falla mediante la aplicación de la técnica mezclas finitas.
- Implementar una herramienta software en MATLAB para la localización de fallas basada en la técnica de clasificación propuesta.

Capítulo 3

Planteamiento del problema

Se estima que el 80 % de las interrupciones ocurren debido a fallas en las redes de distribución [Das, 1998]. Las redes de distribución de un sistema eléctrico de potencia debido a su naturaleza, poseen una compleja estructura que dificulta su rápida revisión. A la vez, la etapa de planeación y operación del mismo se complica a medida que éste crece.

Determinar el origen de una falla de forma rápida, permitiría solventar el problema al reducir el tiempo de atención de las mismas. La mayoría de las fallas son el resultado de corto circuitos. En los sistemas de distribución las fallas por cortocircuito se clasifican en cinco categorías:

- Falla monofásica a tierra.
- Falla bifásica a tierra.
- Falla fase a fase.
- Falla trifásica a tierra.
- Falla trifásica.

Cuando una falla sucede, se produce un cambio dentro del sistema alterando las formas de onda de las señales de tensión y corriente en las fases existentes. El hecho de conocer la característica de tensión y corriente en el sistema de distribución bajo condiciones de falla puede permitir la localización del punto

de falla.

El tema de localización de fallas en los sistemas de distribución comprende un tópico muy importante dentro de las empresas del sector eléctrico. Cuando sucede una falla, se produce una interrupción del servicio y los usuarios se ven afectados parcial o totalmente.

Bajo este escenario, se ha desarrollado múltiples métodos de lclización de fallas a través de procesos analíticos fundamentados sobre tres puntos de vista clásicos.

- **Componenetes de alta frecuencia:** se obtiene información sobre componenetes de alta frecuencia de señales de tensión y corriente medidas bajo condiciones de falla.
- **Fenómenos de ondas viajeras:** relacionados con el comportamiento de las señales de onda de tesión y corriente bajo condiciones de falla.
- **Componentes de frecuencia fundamental:** consiste en el cálculo de la impedancia de falla a través de la medición de las señales de tensión y corriente a frecuencia fundamental. Este cálculo es utilizado para estimar la distancia desde los terminales de la línea hasta el sitio de falla en el sistema.

La inspección visual es un método generalmente empleado en detección de fallas. Las desventaja principal de este método es el alto grado de dependencia de los medios de transporte disponibles, herramientas visuales (binoculares, cámaras de gran alcance, etc), y principalmente de las condiciones climáticas presentes en la zona de inspección. Estos factores transforman este método en un proceso de gran duración durante localización de fallas.

Los algoritmos existentes sobre localización de fallas son poco precisos a la hora de examinar el estado del sistema de distribución debido a la caracterización del sistema, generando inconvenientes de estimaciones múltiples. La incertidumbre en los cálculos se relaciona con aspectos tales como el control,

la operación y la protección [Neimane, 2001]. Estos métodos frecuentemente se combinan con la inspección visual para maximizar la capacidad de localización de fallas. Adicionalmente se implementan dispositivos indicadores distribuidos por tramos en las líneas, que ofrece una guía *in situ* de la ubicación del fenómeno.

Por tanto, desarrollar una herramienta de bajo costo que permita localizar el punto de falla de forma rápida y precisa ayudaría a reducir el tiempo de interrupción del servicio.

El hecho de conocer rápidamente la zona donde se encuentra la falla permite reducir los tiempos de atención del problema. A su vez se mejora los índices FES y DES de la compañía. La reducción de tiempos permite mejorar la planeación y operación del sistema de distribución mediante estrategias prácticas concernientes a la atención de fallas.

Capítulo 4

Antecedentes

Aunque el primer trabajo sobre modelos basados en mezclas de distribuciones fue realizado por el biométrico Karl Pearson en 1894, en los últimos veinte años se han logrado avances considerables en el ajuste de modelos de mezclas finitas. El empleo de modelos de mezclas finitas ha tomado importancia desde su aparición en la monografía sobre Mezclas Finitas (MF) expuesta por McLachlan y Basford en 1988 [McLachlan y Peel, 2000].

En la pasada década, la extensión y el potencial de aplicación de los modelos de MF se han difundido ampliamente. Las mezclas finitas se han aplicado con múltiples propósitos:

- Modelado de la heterogeneidad de una población (Biología).
- Manejo de datos faltantes.
- Estimación de densidades de probabilidad (Estadística).
- Análisis de conglomerados (Clusters).
- Reconocimiento de patrones (Tratamiento de imágenes).
- Obtención de consumos anómalos.

Debido a su flexibilidad, los modelos de mezcla de distribuciones se están explotando como una forma paramétrica de modelar la distribución de poblaciones desconocidas. Con el advenimiento de computadores de alta velocidad

y el rápido desarrollo de técnicas de simulación, se han logrado utilizar los criterios de estimación Bayesianos para el análisis de modelos estadísticos complejos [McLachlan y Peel, 2000]. La estadística computacional, la cual es una colección de técnicas que se enfocan sobre la explotación de los computadores en la creación de nuevas metodologías estadísticas [Martínez *et al*, 2002] permite a los científicos e ingenieros almacenar y procesar una gran cantidad de datos con costos muy bajos. El conjunto de datos que el analista actual maneja tiende a ser muy extenso y multidimensional, lo cual vuelve inadecuado los métodos estadísticos tradicionales.

Tabla 4.1: Comparación entre la estadística tradicional y la estadística computacional

Estadística tradicional	Estadística computacional
Tamaño pequeño y moderado de las muestras	Tamaño muy amplio de muestras
Conjuntos de datos independientes e idénticamente distribuidos	Conjuntos de datos no homogéneos
Pocas dimensiones	Altamente dimensional
Matemáticamente manejable	Numéricamente manejable
Inferencia estadística	Inferencia estructurada
Relaciones de linealidad	Relaciones no lineales

El desarrollo de técnicas complementarias como el algoritmo Expectation-Maximization (EM) y el método de Markov ($MCMC$)[1] , entre otros, ha permitido a los modelos de mezclas finitas una manipulación sencilla de grandes bases de datos.

En el campo de tratamiento de imágenes, la técnica de mezclas finitas se ha utilizado como herramienta de reconocimiento de patrones. En algunas aplicaciones, se maneja el análisis de las texturas como características regionales de las imágenes. Una clasificación de texturas en imágenes, hace referencia a la clasificación de las texturas de una imagen de acuerdo al aspecto de las propiedades extraídas de la imagen propia. La base del análisis de tex-

[1]Markov Chain Monte Carlo, método de estimación de modelos, enfocado sobre técnicas de inferencia Bayesiana [Martínez *et al*, 2002]

turas se relaciona con el procesamiento y análisis de modelos estadísticos y geométricos [Yiming *et al*, 2003]. Esta propuesta se apoya en el hecho de que la distribución de energía en el dominio de la frecuencia identifica una textura. Así, si el espectro de frecuencia de una textura es descompuesto en un número suficiente de sub-bandas, los espectros generados por las diferentes texturas serían lo suficientemente diferentes para obtener una clasificación precisa. Los histogramas generados para cada sub-banda poseen una forma parecida a distribuciones Gaussianas, así que se utiliza el modelo de mezclas para una textura, donde a cada sub-banda se le asigna un componente Gaussiano del modelo de mezcla que lo representa. Para clasificar una imagen cualquiera dentro de los modelos generados, se descompone su espectro en sub-bandas y éstos son evaluados en cada modelo de mezclas. El clasificador seleccionará el tipo de texturas halladas en la imagen de acuerdo a las probabilidades más altas encontradas en los modelos de base.

Otra propuesta ataca el problema mediante la segmentación de la imagen, para usos de compresión de imágenes y optimización de recursos de redes informáticas. La segmentación de imágenes consiste en dividir una imagen en diferentes regiones tal que cada región sea homogénea. Existen tres propuestas muy populares de segmentación de imágenes: las técnicas histogramas de umbrales, el método basado en bordes, y la técnica basada en regiones [Yiming *et al*, 2003]. Mediante el uso de las mezclas normales, aparece una nueva propuesta, utilizando la segmentación de imágenes apoyado en el método de la estimación de la máxima probabilidad. Lo anterior significa que una imagen a color cualquiera, se puede considerar como una mezcla de densidades multivariantes. La segmentación se completa agrupando cada píxel de la imagen dentro de uno de los componentes de la mezcla de acuerdo con la estimación de la máxima probabilidad.

El reconocimiento de patrones utilizando mezclas finitas, también se ha aplicado en sistemas de reconocimiento de voz basados en las funciones densidades de espectro generadas a partir de muestras tomadas a diferentes locutores cuando éstos leen un texto cualquiera. La propuesta genera un cálculo de

probabilidades *a posteriori* basado en un modelo *background*, el cual permite añadir a la base de entrenamiento, el espectro de un nuevo locutor sin necesidad de reestructurar la base completa para un nuevo entrenamiento, lo que ahorra tiempo de procesamiento [Rubio, 2000].

La interpolación y extrapolación espacial es un aspecto muy importante para los sistemas de información geográfica (*GIS*). Los analistas políticos, sociales y de planeación utilizan estos sistemas para clarificar qué pasaría si ocurrieran cambios en algunos aspectos de la población bajo estudio [McLachlan y Peel, 2000]. El uso de mezclas bajo un enfoque Bayesiano permite analizar las relaciones entre variables geo-referenciadas y obtener estimaciones del comportamiento del fenómeno en estudio, acompañado de la estimación de la incertidumbre inherente generada por los parámetros de los modelos aplicados.

Desde el punto de vista financiero, se han utilizado las mezclas de distribuciones dentro de procesos Dirichlet (*DP*), como herramienta de análisis Bayesiano no-paramétrico debido a su flexibilidad y simplicidad. Las mezclas de distribución normales se han utilizado para estimar y predecir densidades que involucran información en tiempo discreto, bajo modelos Gaussianos dinámicos lineales (*DLM's*), que son versiones Bayesianas del popular filtro Kalman [Rodríguez y Horst, 2006]. Esta clase de modelos dinámicos con aproximaciones en tiempo discreto pueden ser fácilmente utilizados en otras áreas, como el modelado de las distribuciones de lluvia, estimaciones de espectros dinámicos y estudios de epidemiología genética.

El problema acerca del diagnóstico del posible lugar de falla en un sistema de distribución está siendo atacado desde diversos puntos de vista. El uso de sistemas informáticos relacionados con inteligencia artificial IA, ha recibido gran atención por parte de los investigadores del tema, implementando métodos como los sistemas expertos, la lógica fuzzy, las redes neuronales y los algoritmos genéticos [Mora, 2003].

Capítulo 5

Fundamento teórico

5.1. Introducción

Las mezclas finitas de distribuciones proveen una aproximación con bases matemáticas al modelado estadístico de una amplia variedad de fenómenos aleatorios. Debido a su utilidad como un método ampliamente flexible de modelado, los modelos de mezclas finitas han tenido un creciente interés a través de los años, desde un enfoque práctico y teórico. La extensión y el potencial de las aplicaciones de las mezclas finitas en las últimas décadas ha sido considerable. Campos del conocimiento tales como la astronomía, la biología, la genética, la medicina, la psiquiatría, la economía, el mercadeo, la ingeniería, entre otras; han aplicado exitosamente modelos estadísticos basados en las mezclas finitas. En estas aplicaciones, los modelos de mezclas finitas han sustentado una variedad de técnicas en muchas áreas de la estadística, incluyendo el análisis de conglomerados, el análisis discriminante, tratamiento de imágenes; en adición con su rol más directo en el análisis de datos, e inferencia que provee modelos descriptivos para las distribuciones.

El primer gran análisis que involucró el uso de modelos de mezclas finitas fue propuesto hace más de cien años por el biométrico Karl Pearson [McLachlan y Peel, 2000]. En 1894, Pearson fijó un modelo basado en la mezcla de dos funciones de densidad de probabilidad normales con diferentes medias μ_1 y μ_2, y varianzas σ_1^2 y σ_2^2; en proporciones π_1 y π_2 de datos proporcionados

por Weldon (1892,1893). Los datos analizados por Pearson, consistían en la medida de la relación entre el frente respecto a la longitud de cuerpo de 1000 cangrejos tomados de la bahía de Naples. Esta medida almacenada en forma de 29 intervalos, mostraba una distribución cercana a la normal, pero de forma asimétrica. En 1893, Weldon había especulado que dicha asimetría en el histograma de los datos era señal de que esta población involucraba dos nuevas subespecies. Bajo la sensación de utilizar un entrenamiento matemático inadecuado, Weldon pidió la ayuda de su colega Karl Pearson.

La aproximación basada en el modelo de mezclas hecho por parte de Pearson (1894), sugirió que existían dos subespecies presentes. Este trabajo fue el primero de dos grandes memorias en una serie de "Contribuciones a la Teoría Matemática de la Evolución" realizada por Stigler (1986) [McLachlan y Peel, 2000]. Sin embargo se destaca el gran esfuerzo matemático realizado por Pearson en la época, para resolver las ecuaciones de noveno grado que determinaban los parámetros que mejor ajustaban el modelo de mezcla a los datos proporcionados.

El desarrollo de nuevas técnicas y de computadores de alta velocidad durante los últimos veinte años, ha facilitado la posibilidad de ajustar los modelos de mezclas finitas a los fenómenos que se pretenden modelar, sobretodo, con datos de más de una dimensión.

Las mezclas finitas ofrecen ventajas sobre la carga computacional que se le exige a un sistema. Existen dos temas a considerar con la mayoría de los métodos de estimación de componentes de densidad de probabilidad. El primero tiene que ver con la carga computacional en términos de la cantidad de información que se debe guardar; y el segundo tema tiene que ver con el esfuerzo computacional requerido para obtener el estimado de la densidad de probabilidad en un punto [Martínez et al, 2002]. En el caso del método de estimación de densidades *kernel*[1], hay que retener todos los datos puntuales,

[1]El *kernel* define una función básica a modo de ventana en análisis de datos generalmente para series de tiempo y frecuencia. En los estimados mediante densidades *kernel*, la distribución de densidad de probabilidad

debido a que el estimado es una suma ponderada de n *kernels* centrados en cada dato puntual. Por tanto se debe calcular el valor del *kernel* n veces. Por supuesto, el tema se complica cuando se manejan estimados multivariados *kernel*, histogramas, y polígonos de frecuencia.

Las mezclas finitas es una técnica para estimar funciones de densidad de probabilidad que puede requerir espacios relativamente pequeños de almacenamiento, y cálculos computacionales pequeños a la hora de evaluar los estimados de las densidades.

5.2. Definiciones básicas

5.2.1. Descripción de datos multivariantes

Describir datos multivariantes supone estudiar cada variable aisladamente, y además las relaciones entre ellas. El objetivo central de la descripción de datos es decidir si los datos son una muestra homogénea de una población o corresponden a una mezcla de poblaciones distintas que deben estudiarse separadamente. El punto central en el análisis de datos es decidir si las propiedades halladas en una muestra pueden generalizarse a la población de la que proviene. Es necesaria la construcción de un modelo del sistema generador de los datos que permita realizar este tipo de extrapolaciones, es decir, hay que suponer una distribución de probabilidad para la variable aleatoria en la población.

Una variable aleatoria es el resultado de observar una característica en un elemento de la población. En el caso multivariado, se analizan variables aleatorias vectoriales. Una variable aleatoria vectorial comprende p características observadas en un elemento de una población. Las variables aleatorias pueden contener información almacenada en forma de cantidades continuas o discretas. Estas últimas pueden ser de carácter cuantitativo (1, 2, 13, etc) o

se genera como la suma ponderada de las funciones *kernel* definidas para cada observación de la muestra. En este tipo de densidades, el problema se centra en la definición del ancho de banda y la función *kernel* a utilizar que generen la densidad de probabilidad requerida.

cualitativo (si, no, negro, alto, etc).

En el caso de observar p variables numéricas en un conjunto de n elementos (observaciones), cada una de estas variables se denomina una variable *escalar* o univariante, y el conjunto de las p variables forma una variable vectorial o multivariante. Los valores de las p variables escalares en cada uno de los n elementos puede representarse en una matriz $\mathbf{X}$, de dimensiones (n x p), la cual recibe el nombre de *matriz de datos* [Peña, 2002]. Cada elemento x_{ij} de la matriz de datos, representa el valor correspondiente a la variable escalar j de la observación i. Por lo tanto cada observación posee información acerca de las p variables aleatorias escalares y se representa por medio de la expresión (5.1).

$$\mathbf{x} = (x_1, x_2, x_3..., x_p) \tag{5.1}$$

Una vez almacenado n observaciones de la muestra en la matriz de datos, se dispone a realizar una descripción de la información almacenada. El punto de inicio de la descripción de datos, parte de establecer una medida de centralización de los mismos. La medida de centralización más utilizada para la describir datos multivariantes es el *vector de medias*, que es un vector de dimensión p cuyas componentes son las medias de cada una de las p variables.

En el estudio univariante, el cálculo de la media para la variable x_j se realiza según la expresión (5.2)

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{5.2}$$

Así, la media consiste en un promedio de las medidas de las n observaciones realizadas para la variable x_j. En el caso multivariante, el promedio de las medidas de cada elemento se realiza en forma vectorial, que define el vector de medias (5.3).

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \dots \\ \overline{x}_p \end{bmatrix} \tag{5.3}$$

El vector de medias se encuentra en el medio de los datos, en el sentido de hacer cero la suma de desviaciones según (5.4).

$$\sum_{i=1}^{n}(x_{ij} - \overline{x}) = 0 \tag{5.4}$$

A partir del vector de medias es posible determinar qué tan homogéneo es el conjunto de datos que se están manejando, una vez se establezcan medidas tales como la variabilidad que dependen de la distancia que existe entre cada observación de la muestra y el vector de medias calculado. La variabilidad respecto a la media, se mide habitualmente por la varianza, o su raíz cuadrada, la desviación estándar (5.5). Esta es definida a través de las desviaciones mediante la distancia entre un punto x_i y la media $d_{ij} = (x_{ij} - \overline{x})^2$ de una variable x:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} d_{ij}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}{n} = 0} \tag{5.5}$$

y su cuadrado es la varianza σ_j^2.

La relación lineal entre dos variables se mide por *la covarianza*. La covarianza entre las demás variables y la variable x_j se calcula por medio de la expresión (5.6).

$$\sigma_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k) \tag{5.6}$$

En el caso multivariante la matriz $\mathbf{S}$ (5.7), se define como la *matriz de varianzas y covarianzas* entre variables de una observación vectorial.

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}}) \tag{5.7}$$

Esta es una matriz cuadrada y simétrica que contiene en la diagonal las varianzas de cada variable y fuera de la diagonal las covarianzas entre variables. Al observar una matriz de varianzas y covarianzas, es posible verificar las características del conjunto de datos que se están analizando. Valores pequeños en la varianza σ_{jj} de la diagonal principal de la matriz, refleja una homogenización significativa de los datos en la variable x_j. De la misma forma, valores muy pequeños de la covarianza σ_{jk} muestra una relación lineal o dependencia entre las variables x_i y x_j. Visto de otra forma, si los valores de las observaciones para la variable x_i tienden a aumentar, la variable x_j también lo hará. Lo anterior sucede cuando se presentan valores positivos de las covarianzas. Cuando la matriz de varianzas y covarianzas presenta valores negativos en los elementos fuera de la diagonal, significa que existen relaciones lineales inversas entre las variables. Lo anterior significa que mientras una de las variables tiende a aumentar su valor, la otra hará exactamente lo contrario.

Es importante saber que el estudio de datos multivariantes, no necesariamente existen relaciones entre las variables de estudio. Esto significa que cada variable funciona de manera independiente y su tendencia no guarda relación alguna con otra en particular. En el caso contrario, puede darse el escenario de que una o más variables sean altamente dependientes de las otras variables de estudio. Estos aspectos permiten evaluar si realmente se están manejando un número óptimo de variables que describan el fenómeno en estudio. El hecho de encontrar una o varias variables linealmente dependientes de otras, hace posible eliminar la información redundante y trabajar con la información que aporta realmente al estudio.

Algebraicamente es posible determinar la existencia de variables linealmente dependientes de otras. Si se analizan los valores propios (*eigenvalores*) de la matriz **S** y se hallan los valores propios iguales a cero, se concluye que la variable asociada a dicho valor propio es una combinación lineal de aquellas variables asociadas con valores propios no nulos. En consecuencia es posible reducir la dimensionalidad del sistema eliminando esta variable (ver apéndice C).

Una vez se manejan variables aleatorias, el concepto de *espacio muestral* entra en juego. Un espacio muestral reúne todos los posibles valores que puede tomar una variable aleatoria vectorial dentro de un subespacio de $\mathbf{R}^n$ con dimensión p. Cuando se define el espacio muestral, es posible empezar a hablar de una función de distribución conjunta.Dada una variable aleatoria vectorial p-dimensional $\mathbf{x} = (x_1, x_2, ..., x_p)$, la función de probabilidad conjunta de una variable aleatoria vectorial $F(\mathbf{x})$ se define en un punto cualquiera $\mathbf{x^0} = (x_1^0, x_1^0, ..., x_p^0)$ mediante (5.8).

$$F(\mathbf{x^0}) = P(\mathbf{x} \leq \mathbf{x^0}) = P(x_1 \leq x_1^0, ..., x_p \leq x_p^0) \tag{5.8}$$

Donde $P(\mathbf{x} \leq \mathbf{x}^0)$ representa la probabilidad de que la variable tome valores menores o iguales al valor considerado $\mathbf{x}^0$. Aunque la función de distribución tiene un gran interés teórico, resulta más cómodo trabajar con funciones de densidad para variables continuas o funciones de probabilidad para variables discretas. La función de densidad conjunta de una variable continua, es la función definida por la función de densidad $f(\mathbf{x})$, que satisface:

$$F(\mathbf{x^0}) = \int_{-\infty}^{\mathbf{x}^0} f(\mathbf{x})d\mathbf{x} = \int_{-\infty}^{\mathbf{x}_1^0} \int_{-\infty}^{\mathbf{x}_2^0} ... \int_{-\infty}^{\mathbf{x}_p^0} f(\mathbf{x})d\mathbf{x} \tag{5.9}$$

La densidad de probabilidad tiene la definición normal de la expresión de densidad: masa por unidad de volumen. Por tanto, la función de densidad conjunta (5.9) debe verificar:

- La densidad es siempre no negativa.
- Si se multiplica la densidad en cada punto por el elemento de volumen en p dimensiones y se integra para todos los puntos con densidad no nula, se obtiene la masa de probabilidad total, que normalmente se estandariza al valor unidad.

Las probabilidades de sucesos definidos como subconjuntos del espacio muestral serán iguales a la masa de probabilidad correspondiente al subconjunto. Estas probabilidades, se calcularán integrando la función de densidad sobre el subconjunto.

5.2.2. Distribución Normal p-dimensional

La distribución Normal o *Gaussiana*, es una de las distribuciones de probabilidad más utilizadas en estadística e ingeniería. La función de densidad de una distribución Normal univariada es definida mediante la expresión (5.10).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{5.10}$$

Donde $-\infty < x < \infty; -\infty < \mu < \infty; \sigma^2 > 0$. La distribución Normal está completamente determinada por sus parámetros (μ y σ^2). Las características principales de las distribuciones Normales son:

- El valor de la función de probabilidad se aproxima a cero en cuanto el valor de x se aleja del centro de la distribución.
- La función de probabilidad está centrada en el valor de la media μ, y el máximo valor de la función ocurre en $x = \mu$.
- La función de probabilidad toma la forma de una campana de forma simétrica alrededor de μ.

En el caso multivariante, un vector $\mathbf{x}$ sigue una distribución Normal p-dimensional (5.11), si su función de densidad es:

$$f(\mathbf{x}) = |\mathbf{V}|^{-\frac{1}{2}}(2\pi)^{-\frac{p}{2}} \exp\left(-\frac{1}{1}2(\mathbf{x}-\mu)'\mathbf{V}^{-1}(\mathbf{x}-\mu)\right) \tag{5.11}$$

Donde $\mathbf{V}$ representa la matriz de covarianzas, y μ el vector de medias de la distribución. La distribución Normal p-dimensional (5.11), mantiene las mismas propiedades expuestas anteriormente, pero se manejan en espacios de dimensiones mayores. Al observar la representación gráfica de diferentes funciones de densidad de distribución Normal, es posible establecer ciertas características de los datos que representan.

En la figura 5.1[2], se distingue tres funciones de densidad Normales con valores diferentes de medias y varianzas. Valores altos de varianza determina una

[2]Fuente: Computacional Statistics Handbook with MATLAB, Martínez Wendy, 2002

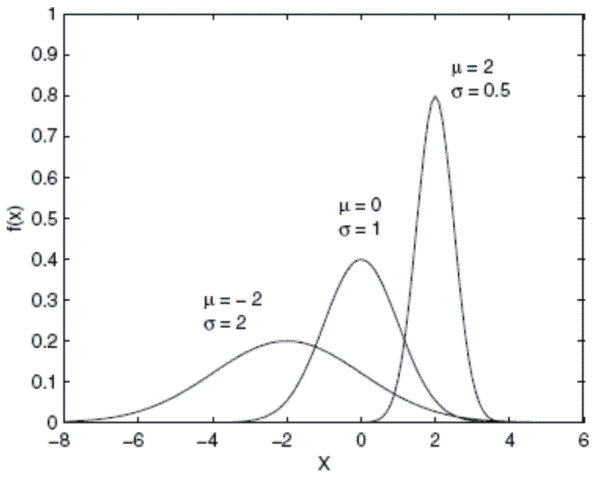

Figura 5.1: Existencia de conglomerados en una sola dimensión

gran dispersión de los datos alrededor de la media. En el caso contrario, la poca dispersión denota una gran tendencia de los datos hacia el valor definido por la media de la distribución.

A menudo una muestra reúne una serie de elementos que aparentemente no poseen una relación entre sí. Es posible que dentro de esta muestra exista la posibilidad de encontrar grupos representativos más pequeños que agrupen los elementos de la muestra, que permitan encontrar características importantes, y no necesariamente estén presentes en todas las observaciones disponibles. En esta instancia es aconsejable realizar un análisis de conglomerados o *clusters*[3] que revele la existencia de subconjuntos ocultos dentro de la muestra.

5.3. Análisis de conglomerados

El análisis de conglomerados tiene por objeto agrupar elementos en grupos homogéneos, en función de las similitudes entre ellos [Peña, 2002]. Este tipo de métodos reciben el nombre de reconocimiento de patrones y estudian tres tipos de problemas:

[3]Se refiere a zonas del espacio muestral, donde coincide una gran cantidad de observaciones de la muestra. El término *racimo*, corresponde a la traducción literal de la expresión cluster.

- Partición de los datos.
  Se tiene una muestra aparentemente heterogénea y se desea dividir en un número de grupos determinado, tal que cada elemento pertenezca a uno, y sólo uno, de los grupos; todos los elementos deben quedar clasificados, y finalmente, cada elemento sea internamente homogéneo. Se utiliza la matriz de datos para realizar este análisis.

- Construcción de jerarquías.
  Se realiza la construcción de una estructura que contenga los elementos de forma jerárquica de acuerdo a su similitud. Los datos se organizan a través de niveles, de manera que los niveles superiores contienen a los inferiores. Este método utiliza la matriz de distancias o similitudes entre objetos.

- Clasificación de variables.
  Realiza un estudio que permita diferenciar las variables en grupos. Este estudio puede orientar hacia la construcción de un modelo más formal que permita reducir la dimensión del espacio muestral.

En este trabajo, el análisis de conglomerados está centrado en el método de partición de datos, y en particular con uno de sus métodos: el algoritmo de k-medias. El algoritmo de k-medias (*k-means*) es un método de partición muy utilizado. El objetivo consiste en hacer una partición de los datos en k grupos diferentes, tal que la suma de cuadrados sea minimizada dentro de los grupos. El algoritmo *k-means* requiere realizar una serie de pasos para su aplicación:

- Selección de k puntos como centros iniciales.
  A partir de la distribución de los datos en el espacio muestral, se escogen k puntos en particular que representen los centros de los grupos a determinar. Esto puede hacerse de forma *a priori*, o estimando de una manera más formal la posible ubicación de los centros de los grupos que se desean formar.

- Cálculo de las distancias Euclídeas de cada elemento a los centro de los k grupos, y asignar cada elemento al grupo cuyo centro esté más próximo.

- Definir un criterio de optimización y comprobar si reasignando algunos de los elementos mejora el criterio. Si no es posible establecer una mejora, terminar con el proceso.
- Minimizar la suma de los cuadrados dentro de los grupos ($SCDG$) (5.12), es el criterio utilizado para optimizar los centros de los grupos dentro de la muestra [Peña, 2002].

$$minSCDG = min \sum_{g=1}^{k} \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ijg} - \overline{x}_{jg})^2 \tag{5.12}$$

Donde x_{ijg} es el valor de la variable j en el elemento i del grupo g, $\overline{x}_{jg}$ es la media de esta variable en el grupo.

Un criterio alternativo es minimizar las distancias al cuadrado entre los puntos y sus centros de grupo, utilizando la norma Euclídea (5.13).

$$min \sum_{g=1}^{k} \sum_{i=1}^{n} (x_{ig} - \overline{x}_g)'(x_{ig} - \overline{x}_g) \tag{5.13}$$

Ambos criterios son idénticos, se busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro. Los resultados del algoritmo dependen de la asignación inicial y el orden de los elementos. Se aconseja repetir el algoritmo con diferentes valores iniciales y permutando los elementos de la muestra. Estos criterios de optimización poseen algunas propiedades como son la no invarianza ante cambios de escala y la producción de grupos aproximadamente esféricos [Peña, 2002].

5.4. Mezclas de distribuciones (Mezclas finitas)

Si existe una muestra aleatoria de tamaño n, donde cada elemento de la muestra representa la observación de una variable aleatoria vectorial p-dimensional $x_i = (x_1, x_2, ..., x_p)$, y se supone que esta muestra posee una función de densidad de probabilidad $f(\mathbf{x})$ en $\mathbf{R}^P$, que puede escribirse mediante:

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f(\mathbf{x}) \tag{5.14}$$

La expresión (5.14) define una función de densidad formada por G funciones de densidad $f_g(\mathbf{x})$, ponderadas por la cantidad π_g. Las G densidades $f_g(\mathbf{x})$, pueden ser funciones de distribuciones de probabilidad Normal, Poisson, exponencial, Gamma, etc. Cada función $f_g(\mathbf{x})$ es una *componente de densidad de la mezcla*. Así la función de densidad $f(\mathbf{x})$, se denomina una *mezcla finita de G-componentes de densidad* y se refiere su correspondiente función de distribución $F(\mathbf{x})$, como una *mezcla finita de G-componentes de distribución*.

Las cantidades π_g, se denominan *coeficientes de mezclado* [McLachlan y Peel, 2000]. Cada coeficiente de mezclado debe cumplir con dos reglas esenciales:

$$0 < \pi_g < 1 \qquad \sum_{g=1}^{G} \pi_g = 1 \tag{5.15}$$

Bajo la formulación dada en (5.15), el número de componentes de la mezcla se considera fijo. Sin embargo, en la mayoría de las aplicaciones, el número de componentes G es desconocido y tiene que ser inferido de los datos disponibles, además de los coeficientes de mezclado y los parámetros que definen las formas de cada componente de densidad.

La forma más fácil de entender los modelos basados en mezclas finitas, es partir del análisis de conglomerados. En el análisis de conglomerados, se busca hallar homogeneidad dentro de una muestra de datos aparentemente heterogénea. Es posible encontrar características similares entre diversas observaciones dentro de una muestra que describe un fenómeno aleatorio, y recoger dentro de un grupo todas las observaciones que posean un mismo rasgo particular. Desde el punto de vista del análisis de conglomerados, un modelo de mezclas finitas describe la posición probabilística dentro de G-componentes de densidad de probabilidad. Cada componente de probabilidad $f_g(\mathbf{x})$, describe la forma cómo se comporta los datos que pertenecen al grupo que representa dentro del espacio muestral. La cantidad π_g, representa la probabilidad *a*

priori de que la observación se localice en el grupo g. La cantidad $\pi_g f_g(\mathbf{x})$, representa la probabilidad *a posteriori* de pertenencia de la observación dentro del grupo g-ésimo. Desde otra perspectiva, los coeficientes de mezclado representan el grado de importancia de cada grupo dentro de la mezcla.

5.4.1. Fundamentos de la estimación máximo verosímil

En el análisis de datos, escoger un modelo que describa de la mejor manera el comportamiento de una población no es tarea fácil. Con base en la información contenida en las observaciones de la muestra se debe definir un modelo el cuál sea capaz de representar y recrear si bien todos, o la mayoría de los aspectos que caracterizan a la población bajo estudio. El método de máxima verosimilitud expuesto por Fisher (1922) [Peña, 2002], escoge el estimador de los parámetros, aquel que hace máxima la probabilidad de que el modelo a estimar genere la muestra observada. Por lo general las muestras son datos parciales que representan una población. Bajo estas condiciones los parámetros de las funciones de densidad, en el caso de las distribuciones Normales: $\theta = (\mu, \mathbf{V})$, son desconocidos.

Un estimador $\hat{\theta}$, es una función que busca representar los parámetros poblacionales desconocidos de la mejor manera posible. Con la ayuda de los datos muestrales $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, es posible llegar a un valor en particular del estimador $\hat{\theta}$, de modo que brinde una conjetura razonable acerca de un parámetro poblacional real.

Lo anterior se denomina una estimación puntual $\hat{\theta}_i$. A la hora de estimar el valor de los parámetros desconocidos, se dispone de varios estimadores que representan diferentes modelos. En teoría un buen estimador debe estar lo más cercano posible al valor del vector θ, con parámetros desconocidos. Analizando la distribución generada por un estimador $\hat{\theta}$ dado, si $E(\hat{\theta}) = \theta$[4] , entonces el estimador es una aproximación muy buena de θ (estimador inses-

[4]El término define la esperanza matemática del estimador $\hat{\theta}_i$. Según la expresión, el valor esperado del estimador debe ser equivalente a θ, el cual genera la función de distribución de los datos de la población.

gado)[Peña, 2002]. En el caso contrario, existe una diferencia entre $E(\hat{\theta})$ y θ lo cual se denomina *sesgo*. El sesgo define la diferencia cuantitativa entre θ en relación con $\hat{\theta}$.

Si se tiene una muestra aleatoria simple de n elementos de una variable aleatoria p-dimensional $\mathbf{x}$, con función de densidad $f(\mathbf{x}|\theta)$, donde $\theta = (\theta_1, \theta_2, ..., \theta_p)'$ es un vector de parámetros con dimensión $r < pn$. La función de densidad conjunta (5.16) para la muestra de datos $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, por la dependencia de las observaciones estará definida por:

$$f(\mathbf{X}|\theta) = \prod_{g=1}^{n} f(\mathbf{x}_i|\theta) \tag{5.16}$$

Cuando se conoce θ, la función (5.16) determina la probabilidad de aparición de cada muestra. En el problema de estimaciones, se dispone de la muestra, pero no se conoce θ. Si se considera θ dentro de la expresión de la densidad conjunta como una variable y se define (5.16) para los datos observados, se obtiene una función denominada *función de verosimilitud* $l(\theta|\mathbf{X})$, o $l(\theta)$.

$$l(\theta|\mathbf{X}) = \prod_{g=1}^{n} f(\mathbf{x}_i|\theta) \qquad \mathbf{X}\ fijo, \qquad \theta\ variable \tag{5.17}$$

El estimador de máxima verosimilitud MV, es aquel que hace máxima la probabilidad de aparición de los valores muestrales observados, y se halla calculando el valor máximo de la función de verosimilitud (5.17). Se puede obtener un máximo, si se supone que la función de verosimilitud es diferenciable y que dicho máximo no ocurre en un extremo de su dominio de definición, con lo cual se puede resolver el sistema de ecuaciones:

$$\frac{\partial l(\theta)}{\partial \theta_1} = 0 \qquad , ..., \qquad \frac{\partial l(\theta)}{\partial \theta_r} = 0 \tag{5.18}$$

El vector $\hat{\theta}_i$ que satisface este sistema de ecuaciones (5.18) corresponde a un máximo si la matriz Hessiana de segundas derivadas $\mathbf{H}$, evaluada en $\hat{\theta}_i$, es definida negativa según (5.19).

$$\mathbf{H}(\hat{\theta}) = \left(\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right)_{\theta=\hat{\theta}} \qquad se\ define\ negativa \tag{5.19}$$

En ese caso es el estimador MV de θ. En la práctica suele ser más fácil obtener el máximo del logaritmo de la función de verosimilitud, llamada *función de soporte* (5.20).

$$L(\theta) = \ln l(\theta) \tag{5.20}$$

El logaritmo por ser una función monótona, las expresiones en (5.20) tienen el mismo máximo, pero trabajar con la función soporte tiene tres ventajas fundamentales. Primero se pasa de un producto de densidades a la suma de sus logaritmos, con una expresión mucho más simple, con la cual se obtiene el máximo muy cómodamente. En segundo lugar, las constantes multiplicativas de la función de densidad, se hacen aditivas y desaparecen al derivar; con lo que la derivada del soporte toma siempre la misma forma y no depende de constantes arbitrarias. Por último, el doble de la función soporte con signo opuesto, proporciona un método general para juzgar el ajuste del modelo a los datos, que se denomina *desviación* [Peña, 2002]:

$$D(\theta) = -2L(\theta) \tag{5.21}$$

La desviación (5.21) mide la discrepancia entre el modelo y los datos. Cuanto mayor sea el soporte, mayor es la concordancia entre el valor del parámetro y los datos. En condiciones muy generales respecto al modelo de distribución de probabilidad, el método de máxima verosimilitud proporciona estimadores que son:

- Asintóticamente centrados, con distribución asintóticamente normal.
- Asintóticamente de varianza mínima (son eficientes).
- Si existe un estadístico suficiente para el parámetro, el estimador MV es suficiente.

- Es invariante: si $\hat{\theta}$ es el estimador MV de θ, y $g(\theta)$ es una función cualquiera del vector de parámetros, entonces, en condiciones bastante generales, $g(\hat{\theta})$ es el estimador MV de $g(\theta)$.

5.4.2. Estimación de mezclas finitas normales

Si se supone que los datos de la muestra fueron generados por una mezcla de distribuciones normales (5.22), la forma para estimar los parámetros de las distribuciones implicadas y las probabilidades a priori de pertenencia de cada dato a cada una de las componentes de la mezcla empieza calculando la función de verosimilitud de la distribución.

$$f(\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i) \tag{5.22}$$

La función de verosimilitud asociada a la expresión (5.22) será:

$$l(\theta|\mathbf{X}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i) \tag{5.23}$$

La expresión (5.23) se puede escribir como la suma de los términos correspondientes a todas las posibles clasificaciones de las n observaciones entre los G grupos existentes (G^n). Por lo tanto la función soporte de la muestra será:

$$L(\theta|\mathbf{X}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i) = \sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i) \tag{5.24}$$

Si cada función de densidad se considera una función normal k-dimensional con vector de medias μ_g y matriz de covarianzas V_g, de manera que $\theta = (\pi_1, ..., \pi_G, \mu_1, ..., \mu_G, V_1, ..., V_G)$. Al sustituir todas las densidades por sus expresiones, la función soporte (5.24) toma la forma:

$$L(\theta|\mathbf{X}) = \sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g |\mathbf{V}_g|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \exp\left(-\frac{(\mathbf{x}_i - \mu_g)' \mathbf{V}_g^{-1} (\mathbf{x}_i - \mu_g)}{2}\right) \tag{5.25}$$

Si $\mu_g = \mathbf{x}_i$, la estimación de $\mathbf{V}_g$ es cero y si $\pi_g \neq 0$, el cociente $\pi_g |\mathbf{V}_g|^{-\frac{1}{2}}$ tendería a infinito y también lo haría la función soporte. Por tanto la función

(5.25) tendría una gran cantidad de máximos, relacionados con soluciones donde cada densidad viene determinada por una observación. Para evitar caer en tales singularidades, se supone que, como mínimo existen p-observaciones en cada distribución, para tratar de hallar un máximo local de la función que brinde un estimador lo suficientemente consistente de los parámetros.

Para maximizar (5.25) con relación a las probabilidades π_g, hay que considerar que $\sum_{g=1}^{G} \pi_g = 1$. Bajo esta restricción, aplicado a un multiplicador de Lagrange, la función a maximizar es:

$$L(\theta|\mathbf{X}) = \sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i) - \lambda(\sum_{g=1}^{G} \pi_g - 1) \tag{5.26}$$

Derivando (5.26) respecto a las probabilidades *a priori*:

$$\frac{\partial L(\theta|\mathbf{X})}{\partial \pi_g} = \sum_{i=1}^{n} \frac{f_g(\mathbf{x}_i)}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} - \lambda \tag{5.27}$$

Multiplicando (5.27) por π_g, se puede definir la expresión (5.28). Si se supone que $\pi_g \neq 0$, de lo contrario el modelo g es redundante.

$$\lambda \pi_g = \sum_{i=1}^{n} \pi_{ig} \tag{5.28}$$

Donde π_{ig} se denomina:

$$\pi_{ig} = \frac{\pi_g f_g(\mathbf{x}_i)}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} \tag{5.29}$$

Los coeficientes π_{ig} en la expresión (5.29), representan la probabilidad de que una vez observado el dato $\mathbf{x}_i$, éste haya sido generado por la distribución Normal $f_g(\mathbf{x})$. Estas probabilidades se denominan a posteriori y son calculadas por el teorema de Bayes. Antes de observar $\mathbf{x}_i$, la probabilidad de que cualquier observación, y en particular $\mathbf{x}_i$, venga de la clase g es π_g. Esta probabilidad se modifica después de observar $\mathbf{x}_i$ en función de lo compatible que sea este valor con el modelo g. Dicha compatibilidad se mide por $f_g(\mathbf{x}_i)$. A medida que este valor sea relativamente alto, aumentará la posibilidad de que

venga del modelo g.

El valor de λ se determina sumando (5.28) para todos los grupos.

$$\lambda = \sum_{i=1}^{n} \sum_{g=1}^{G} \pi_{ig} = n \tag{5.30}$$

Sustituyendo (5.30) en (5.28), las ecuaciones para estimar las probabilidades *a priori* son:

$$\hat{\pi_g} = \frac{1}{n} \sum_{i=1}^{n} \pi_{ig} \tag{5.31}$$

La expresión (5.31) proporciona las probabilidades *a priori* como un promedio de las probabilidades *a posteriori*.

Al derivar la función soporte (5.26) respecto de las medias, es posible calcular las estimaciones de los parámetros de las distribuciones.

$$\frac{\partial L(\theta|\mathbf{X})}{\partial \mu_g} = \sum_{i=1}^{n} \frac{\pi_g f_g(\mathbf{x}_i)\ \mathbf{V}^{-1}(\mathbf{x}_i - \mu_g)}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} = 0, \qquad g = 1, ..., G \tag{5.32}$$

La expresión (5.32) se puede representar como:

$$\hat{\mu}_g = \sum_{i=1}^{n} \frac{\pi_{ig}}{\sum_{i=1}^{n} \pi_{ig}} \mathbf{x}_i \tag{5.33}$$

La media (5.33) de cada distribución se estima como una media ponderada de todas las observaciones con pesos $w_i = \sum_{i=1}^{n} \frac{\pi_{ig}}{\sum_{i=1}^{n} \pi_{ig}}$ donde $w_i \geq 0$, y $\sum_{i=1}^{n} w_{ig} = 1$. Los pesos w_{ig}, representan la probabilidad relativa de que la observación i pertenezca a la población g. De manera similar, derivando (5.26) respecto a $\mathbf{V}_g$ se puede obtener, la varianza (5.34) para cada distribución.

$$\hat{\mathbf{V}}_g = \sum_{i=1}^{n} \frac{\pi_{ig}}{\sum_{i=1}^{n} \pi_{ig}} (\mathbf{x}_i - \hat{\mu}_g)(\mathbf{x}_i - \hat{\mu}_g)' \tag{5.34}$$

Si bien las ecuaciones (5.31), (5.33) y (5.34) proporcionan los estimadores de las componentes de la mezcla, es necesario calcular primero las probabilidades

π_{ig}. Para calcular estas probabilidades con (5.29), se necesitan los parámetros del modelo. Es necesario fijar unas condiciones iniciales que proporcionen los parámetros de arranque para los cálculos de los estimadores del modelo, como primer paso a seguir.

El empleo de las ecuaciones de forma iterativa permite obtener estimadores que describan lo mejor posible a los datos de la muestra. Ésta solución es la que se obtiene con el algoritmo *EM*.

5.5. El algoritmo *EM*

El algoritmo *EM* (*Expectation-Maximitation*) es un método para optimizar las funciones de probabilidad. La metodología del *EM* es ahora una herramienta estándar para los estadísticos y se utiliza en muchas aplicaciones. El problema es determinar los parámetros del modelo que describa a una población en particular. Por lo general sólo se dispone de una cantidad de datos parciales agrupados en la muestra. Si bien la muestra es una parte representativa de la población, a menudo existen una cantidad de datos ausentes que dificultan una estimación precisa de los estadísticos necesarios para generar los datos de la población.

Inicialmente se encuentran observaciones con sus datos completos, pero las últimas observaciones presentan parcial o totalmente la ausencia de valores en sus variables. Por tanto se dispone de una muestra $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, \mathbf{x'}_{n+1}, \mathbf{x'}_{n+2}, ..., \mathbf{x'}_N)$, donde los $\mathbf{x'}_i$ presentan ausencias. Por otro lado existen observaciones con la totalidad de sus valores $\mathbf{x}_i$ y otras simplemente carecen de valores en sus variables $\mathbf{z}_i$. Estas últimas están distribuidas en la muestra sin un orden en particular. Para este caso es necesario realizar un reordenamiento de forma tal que la muestra presente la siguiente forma: $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, \mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m)$.

Gracias al algoritmo *EM* es posible utilizar los datos en condiciones de valores ausentes junto con el criterio de la máxima verosimilitud para obtener un

modelo apropiado a los datos presentes. El algoritmo se divide en dos partes.

El primer paso denominado estimación (E), trabaja con la función soporte $L(\mathbf{Y}, \mathbf{Z}|\hat{\theta})$ de la muestra, hallando la esperanza (5.35) de las funciones de los datos ausentes $\mathbf{Z}$ a partir de determinar la distribución de $\mathbf{Z}$ con los valores iniciales y los datos observados $\mathbf{Y}$ [Peña,2002].

$$L(\theta|\mathbf{Y}) = E(l(\theta|\mathbf{Y}, \mathbf{Z}) \tag{5.35}$$

El segundo paso, se denomina maximización (M). En este paso se maximiza la función de soporte respecto a θ con el fin de conseguir los estimadores MV a partir de la sustitución de las observaciones faltantes por estimaciones de sus valores [Martínez *et al*, 2002].

El algoritmo EM comienza con una estimación inicial de los parámetros $\hat{\theta}_0$. En el paso$E$, se calcula el valor esperado de las observaciones ausentes en la verosimilitud completa condicionando a los parámetros iniciales y a los datos observados. En la etapa M se maximiza esta función respecto de los parámetros $\hat{\theta}_0$. Las ecuaciones (5.31),(5.33) y (5.34) descritas anteriormente, son las directamente implicadas durante la ejecución del algoritmo.

Los pasos son los siguientes:

1. Determinar el número de términos o componentes de densidad G dentro de la muestra.

2. Determinar un supuesto inicial de los valores de los parámetros. Estos son, los coeficientes de mezclado, las medias y las matrices de covarianzas para cada función de densidad Normal.

3. Para cada elemento $\mathbf{x}_j$, se calcula la probabilidad posterior utilizando la expresión (5.29).

4. Se actualizan los coeficientes de mezclado, las medias y las matrices de covarianzas para cada componente utilizando las ecuaciones (5.31),(5.33) y (5.34).

5. Se repiten los pasos 3 y 4 hasta que el estimado converja.

Sea $\hat{\theta}_{i+1}$ el estimador obtenido en el paso M, se retorna al paso E de manera iterativa hasta obtener la convergencia deseada, es decir, hasta que $|\hat{\theta}_{i+1} - \hat{\theta}_i|$ sea lo suficientemente pequeño [Martínez *et al*, 2002].

5.6. Evaluación de número de componentes en los modelos de mezclas de distribución

El método de máxima verosimilitud supone que la forma del modelo es conocida y sólo falta estimar los parámetros. Cuando no es así, debe aplicarse con cuidado con el propósito de obtener modelos que representen la muestra de la mejor manera.

Si se supone que existe una muestra, y paralelo a ello, se dispone de varios modelos con número diferente de componentes que representan la muestra existente. Al utilizar el método de máxima verosimilitud, éste siempre dará mayor soporte al modelo que tenga más parámetros, pues su valor sólo puede aumentar al introducir mayor cantidad de parámetros para explicar los datos [Peña, 2002]. Sin embargo no necesariamente el modelo con mayor número de parámetros es aquel que mejor describe los datos de la muestra.

Este aspecto del método de máxima verosimilitud fue percibido por Fisher (1936), que propuso el método para estimar parámetros de un modelo, indicando sus limitaciones para comparar modelos distintos. Habitualmente se realiza un contraste entre las verosimilitudes de los modelos (5.36), eligiendo al modelo M_i frente al M_j. Lo anterior se realiza mediante la comparación de las desviaciones.

$$\lambda = 2(L(M_j) - L(M_i)) = D(M_j) - D(M_i) \tag{5.36}$$

5.6.1. El criterio de Akaike *AIC*

Hirotsugu Akaike (1974) propuso un enfoque alternativo para resolver el problema de seleccionar modelos suponiendo que el objetivo es hacer predicciones tan precisas como sea posible. Si $f(y|M_i)$ es la densidad de una nueva observación bajo el modelo M_i, $f(y)$ es la verdadera función de densidad (la verdadera función de densidad puede o no ser uno de los modelos considerados en la evaluación), y se desea seleccionar el modelo de manera que $f(y|M_i)$ sea tan próxima como sea posible a $f(y)$. Una forma de medir las distancias entre las dos funciones de densidad es mediante la divergencia de Kulback-Liebler (1951) [McLachlan y Peel, 2000], que se calcula:

$$KL(f(y|M_i), f(y)) = \int \log \frac{f(y|M_i)}{f(y)} f(y)dy \tag{5.37}$$

Cuando los valores de ambas funciones en (5.37) son similares, la diferencia de los logaritmos equivale a la diferencia relativa:

$$\log \frac{f(y|M_i)}{f(y)} = \log \left(1 + \frac{f(y|M_i) - f(y)}{f(y)}\right) \cong \frac{f(y|M_i) - f(y)}{f(y)} \tag{5.38}$$

Cuando las diferencias son grandes, el logaritmo es la mejor medida de discrepancia que la diferencia relativa. Las discrepancias se promedian respecto a la verdadera distribución de la observación y la medida (5.38) siempre será positiva. Esta distancia se puede minimizar la distancia entre la verdadera distribución y $f(y|M_i)$, haciendo el primer término lo más pequeño posible. Esto equivale a minimizar la expresión:

$$-2L(M_i) + 2p_i = D(M_i) + 2p_i \tag{5.39}$$

Donde p_i es el número de parámetros de modelo M_i. El criterio es disminuir la suma de la desviación del modelo, que disminuirá si se introducen más parámetros; más el doble del número de parámetros en el modelo, que tiende a corregir este efecto. La expresión (5.39), se conoce como criterio de Akaike.

5.6.2. Criterio de información Bayesiano *BIC*

El criterio de información Bayesiano (*Bayesian Information Criterion*) propuesto por Schwarz (1978) [McLachlan y Peel, 2000], supone abordar el problema desde un enfoque bayesiano. Al considerar los modelos como posibles hipótesis sobre lo datos, se calcularán sus probabilidades *a posteriori* y se escogerá el modelo con máxima probabilidad *a posteriori*. Estas probabilidades viene dadas por:

$$P(M_i|\mathbf{X}) = \frac{f(\mathbf{X}|M_i)}{f(\mathbf{X})}P(M_i) \qquad i = 1, 2, ..., m \tag{5.40}$$

Donde (5.40) es la probabilidad *a priori* del modelo j. Esta ecuación indica cómo se pasa de la probabilidad *a priori* a la probabilidad *a posteriori* para cada modelo: se calcula la verosimilitud marginal de los datos para ese modelo $f(\mathbf{X}|M_i)$, donde el nombre de marginal se debe a la independencia de esta función de los valores de los parámetros y se compara con la verosimilitud marginal promedio de todos los modelos, $f(\mathbf{X})$. Al final se obtiene una expresión similar al criterio *AIC*:

$$-2L(\hat{\theta}|\mathbf{X}) + 2p_i \log n \tag{5.41}$$

Donde n es el número de elementos de la muestra. Schwarz propuso escoger el modelo que conduzca a un valor mínimo de esta cantidad. Este criterio pondera la desviación del modelo, con el número de parámetros [Peña, 2002]. Si se introducen más parámetros en el modelo, mejorará el ajuste, así el soporte aumenta o disminuye la desviación. Este efecto queda compensado por el número de parámetros que aparecen en $p_i \log n$. La ecuación (5.41), se conoce como el criterio de información bayesiana (*BIC*).

5.6.3. Criterio de Clasificación de Probabilidad Integrada *ICL*

El criterio de clasificación de probabilidad integrada fue propuesto por Biernacki (1998) [McLachlan y Peel, 2000], como un intento de superar los defectos de los criterios *BIC* y *CLC* (*Criterio de Clasificación de Probabilidad*). La expresión (5.42) define el criterio *ICL*.

$$-2\log L(\hat{\theta}_j) + 2EN(\hat{\pi}) + d\log n \tag{5.42}$$

Donde d es el número de parámetros desconocidos y n es el número de elementos de la muestra. La expresión $EN(\hat{\pi})$, se denomina la *entropía* de la matriz de clasificación difusa (5.43).

$$EN(\hat{\pi}) = \sum_{g=1}^{G}\sum_{i=1}^{n} \pi_{ig}\mathbf{x}_i \tag{5.43}$$

Si se observan los criterios de evaluación descritos anteriormente, se distingue una parte común entre ellos. En todos aparece la expresión de la desviación, la cual describe lo acertado del modelo. A medida que aumenta el número de parámetros en el modelo, el soporte aumenta, con lo que disminuye la desviación. La segunda parte de las expresiones en cada criterio se denominan regularmente como penalizaciones de los modelos en estudio [McLachlan y Peel, 2000]. A medida que aumenta el número de parámetros en los modelos, el valor de la penalización también lo hará. Por tanto, escoger modelos con números muy grandes de parámetros supone mejorar la desviación en éstos, pero esta selección es altamente penalizada por el hecho de escoger modelos demasiado complejos.

5.7. Clasificación de datos de una población utilizando mezclas de distribuciones

Una vez realizada la estimación de modelos utilizando las mezclas de distribuciones, es posible utilizar el modelo seleccionado para determinar la procedencia de los datos de una población. La clasificación de los datos de una población se realiza mediante el análisis discriminante. Este tipo de análisis supone realizar lo que se conoce como reconocimiento de patrones (*Pattern Recognition*)[Martínez *et al*, 2002]. El reconocimiento de patrones permite establecer las características de una población con base en la aplicación de un modelo estadístico que describa un comportamiento en particular.

Si se dispone de un conjunto amplio de elementos que pueden venir de dos o más poblaciones distintas, es posible establecer la probabilidad de que cada elemento provenga de alguna de las poblaciones conocidas en el modelo. Si se conoce las probabilidades *a priori* de que el elemento venga de cada una de las poblaciones, además de los parámetros que caracterizan a cada población, entonces tenemos una distribución mezclada de acuerdo a la expresión (5.44).

$$f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x}) + ... + \pi_G f_G(\mathbf{x}) \tag{5.44}$$

El cálculo de probabilidad *a posteriori* de que el elemento haya sido generado por cada una de las poblaciones, puede hallarse mediante el teorema de Bayes. Para el caso del elemento $\mathbf{x}_0$ su probabilidad *a posteriori* de pertenecer a la distribución i de la mezcla está dada por la expresión (5.45).

$$P(1|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|1)\pi_1}{P(\mathbf{x}_0|1)\pi_1 + P(\mathbf{x}_0|1)\pi_2 + ... + P(\mathbf{x}_0|1)\pi_G} \tag{5.45}$$

Se clasificará la observación $\mathbf{x}_0$ en la población más probable *a posteriori*.

Todo el contexto presentado anteriormente proporciona las herramientas necesarias para aprovechar la información contenida en los datos de la muestra. Los valores arrojados en los resultados no muestran la solución por sí mismos. Es necesario sobreponer esta información, a los fundamentos concernientes al campo de aplicación de todas estas técnicas. Sólo así toda la información generada tendrá un verdadero significado.

Capítulo 6

Desarrollo de la metodología

6.1. Introducción

Es posible resolver el problema de la localización de fallas, si se conoce con exactitud el comportamiento de los sistemas cuando estos se encuentran sometidos a perturbaciones durante cada evento. La respuesta de los sistemas puede verse reflejada en las señales de tensión y corriente registradas durante la presencia de fallas.

El análisis desde el punto de vista estadístico sobre la forma del comportamiento de un sistema, supone obtener un historial o base de datos adecuada en la cual se reúna información suficiente del sistema sometido a condiciones de falla.

6.2. Características del sistema de potencia de prueba

Antes del desarrollo y análisis de las herramientas, se seleccionó un sistema prototipo en el cual se pueda representar las características de los sistemas de distribución, esto es, exista topología radial, secciones de línea, cargas (monofásicas y trifásicas), así como derivaciones. En el sistema prototipo representado en la figura 6.1, existen 21 nodos, con los equipos de medida ubicados en la cabecera del circuito.

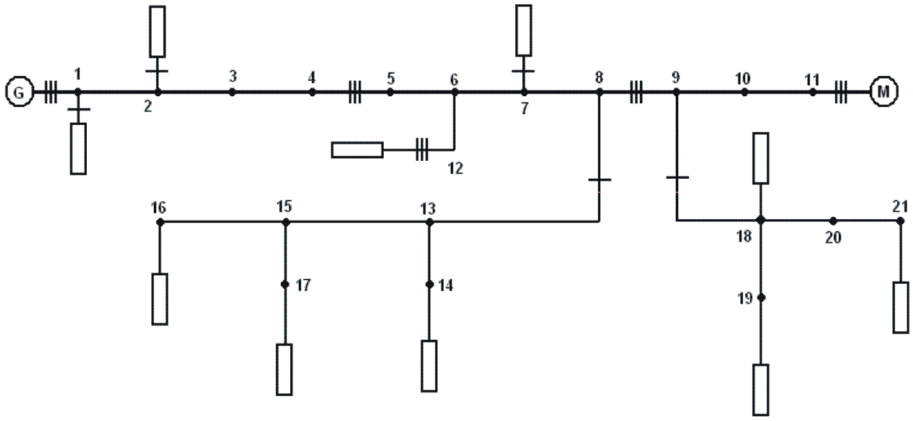

Figura 6.1: Sistema de distribución prototipo

6.3. Simulación del sistema de potencia en condiciones de falla

Las fallas más comunes son el resultado de cortocircuitos. Por lo que se tuvieron presentes las fallas del tipo monofásica a tierra, bifásica, bifásica a tierra, trifásica y trifásica a tierra; cada una con diferentes valores de resistencia de falla. El sistema propuesto se implementó en el software ATP/EMTP con el fin de realizar simulaciones del sistema bajo las condiciones de falla mencionadas anteriormente. A través de las simulaciones se construyó un banco de señales con registros de tensión y corriente, para fallas ocurridas en cada punto del sistema. Los datos obtenidos fueron distribuidos en dos grupos, para realizar los procesos de entrenamiento (E) y validación (V), tal como se muestra en la tabla 6.1.

6.4. Manejo de los datos obtenidos por simulación

Los datos obtenidos por simulación representan la población en la cual se analizará el comportamiento del sistema. Una parte de los datos se seleccionó como muestra, para efectos de análisis, desarrollo y entrenamiento de los algoritmos responsables de generar los modelos. Los datos de entrenamiento tendrán un carácter informativo por lo que sus características no serán des-

Tabla 6.1: Consolidado de simulaciones en el sistema de distribución prototipo

Tipo de falla	No. simulaciones			Proceso	
	Fase A	Fase B	Fase A	E	V
Monofásica	132	187	176	315	180
Bifásica LL	132	132	132	252	144
Bifásica LLT	132	132	132	252	144
Trifásica		132		84	48
Trifásica a tierra		132		84	48
Total		**1551**		**1323**	**564**

conocidas para el analista. Lo anterior significa que previamente se conoce en gran parte, las características de cada observación de entrenamiento tales como: tipo de falla, valor de resistencia de falla, ubicación dentro del sistema, tiempo de despeje de la falla, etc. La razón para tratar estos datos como conocidos se relaciona con la capacidad de asociar los valores de los parámetros producidos por las herramientas de análisis estadístico, con los fenómenos registrados por las señales de tensión y corriente recopiladas. Las observaciones restantes fueron seleccionadas para efectos de evaluación y prueba de los modelos generados (ver capítulo 7). Estos datos están dispuestos para validar la capacidad de los modelos generados para describir el sistema en estudio, y estimar correctamente el tipo de información suministrada a la hora de caracterizar una falla. En pocas palabras, estos datos sirven para evaluar la eficiencia de los clasificadores generados.

La información necesaria para generar los modelos estadísticos se agrupa en forma de observaciones multivariantes las cuales recopilan la información del comportamiento del sistema sometido a cada una de las condiciones de falla, por medio de variables estadísticas denominadas *descriptores*[1] . Escoger el tipo de descriptores y cuántos de ellos se manejan, determina la calidad de las estimaciones esperadas y el grado de complejidad de los modelos a obtener.

[1]Un descriptor define una característica particular de las observaciones tomadas de un evento aleatorio.

La utilización de una gran cantidad de descriptores supone obtener modelos mucho más precisos, en espacios vectoriales con un número mayor de dimensiones, lo que dificulta la estimación de los modelos [Peña, 2002]. De igual forma, es posible obtener modelos estadísticos muy precisos con cantidades menores de descriptores adecuadamente seleccionados, en comparación con modelos que necesitan una gran cantidad de descriptores, alcanzando precisiones similares.

La propuesta para generar los modelos estadísticos que permitan caracterizar el sistema de distribución en estudio fue:

- Realizar una selección de descriptores adecuados para su tratamiento estadístico.
- Utilizar técnicas de exploración visual y reconocimiento de patrones, para encontrar rasgos característicos relacionados con el tipo de falla, y su ubicación dentro del sistema (Se tomó cada tipo de falla por separado, y una a una, se analizaron para encontrar los patrones asociados a cada tipo de falla).
- Aplicar la técnica de mezclas finitas y sus herramientas, para generar modelos que caractericen el comportamiento del sistema bajo falla.
- Determinar cuál es la mejor forma de generar modelos que brinden óptimos resultados.

6.4.1. Definición de descriptores

Al ocurrir una falla, se presenta una variación de las señales de tensión y corriente existentes en un sistema. Una falla se conoce como cualquier cambio en un sistema que evita la correcta operación del mismo [Mora, 2003]. En algunos sistemas expertos, las señales de tensión y corriente se han dividido en sectores representativos de la forma de onda cuando sucede una falla. La primera aproximación al análisis de los datos generados en la etapa de simulación comprendió la obtención de estos valores para ser utilizados como

descriptores de la base de datos. Los descriptores generado de acuerdo a la variación de la señal de tensión y corriente son los siguientes:

- **Pendiente de ascenso del pico de corriente** ($\mathbf{ps_I}$).
  Valor absoluto de la pendiente de ascenso de corriente. Este se valor se mide en el intervalo I (figura 6.2). Desde el momento en que el valor de corriente aumenta 10 % del valor nominal y el tiempo en el cuál se alcanza el 5 % por debajo del valor h_I.

- **Magnitud máxima del pico de corriente por fase** ($\mathbf{h_I}$).
  Es el máximo valor RMS por fase, alcanzado por la señal de corriente durante el evento de falla (figura 6.2).

- **Pendiente de bajada del pico de corriente** ($\mathbf{pb_I}$).
  Valor absoluto de la pendiente de descenso de corriente. Este se valor se mide en el intervalo III (figura 6.2). Desde el momento en que el valor de corriente alcanza el 5 % por debajo del valor h_I y el tiempo en el cuál se mide el 10 % por encima del valor nominal de corriente.

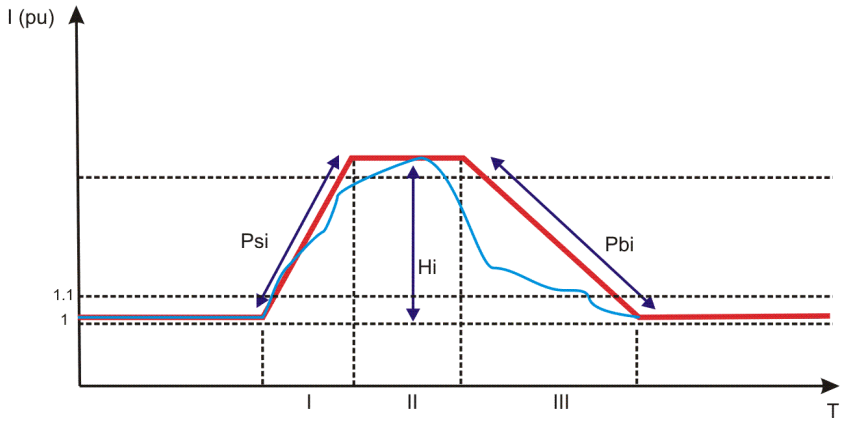

Figura 6.2: Representación gráfica de la señal de corriente durante una falla

- **Pendiente de bajada del hueco de tensión** ($\mathbf{pb_V}$).
  Valor absoluto de la pendiente de caída de tensión. Este se valor se mide en el intervalo I (figura 6.3). Desde el momento en que el valor de tensión disminuye 10 % del valor nominal y el tiempo en el cuál se alcanza el 5 % por encima del valor h_V.

- **Magnitud máxima de la caída de tensión**[2] **por fase** ($\mathbf{h_V}$).
  Es el mínimo valor RMS por fase, alcanzado por la señal de tensión durante el evento de falla (figura 6.3).

- **Pendiente de subida del hueco de tensión** ($\mathbf{ps_V}$).
  Valor absoluto de la pendiente de ascenso de tensión. Este valor se mide en el intervalo III (figura 6.3). Desde el momento en que el valor de corriente aumenta 5 % por encima del valor h_V y el tiempo en el cuál se alcanza el 90 % del valor nominal de tensión.

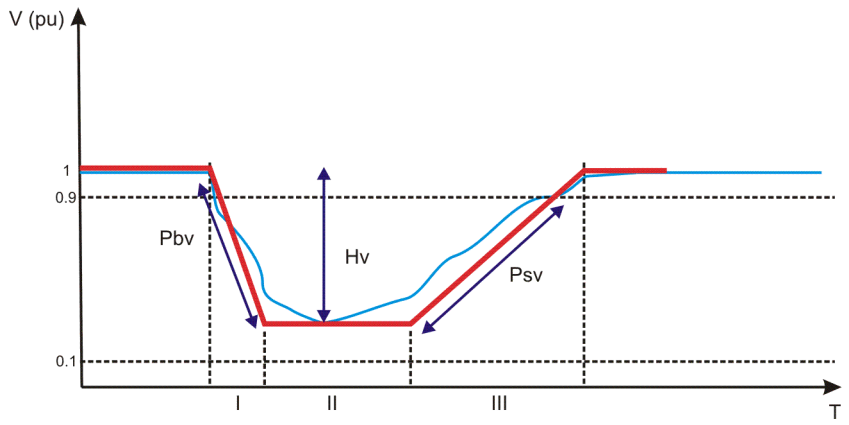

Figura 6.3: Representación gráfica de la señal de tensión durante una falla

[2]Una caída de tensión, se define como la reducción entre 0,9 y 0,1p.u. del valor nominal RMS de tensión, durante un período entre 8,33ms y 1 minuto, a frecuencia industrial [IEEEstd1159-1995,1995].

6.4.2. Selección de descriptores utilizando minería de datos

La exploración visual de los datos multivariables se convierte una herramienta poderosa para el análisis de la información. La exploración visual permite estudiar las formas de las distribuciones de los datos antes de aplicar técnicas analíticas. Durante la etapa de exploración de datos, se observó que la información sobre los tiempos de duración de la falla, y las pendientes, pb_I y ps_V aportan información relativa, debido a su relación con las características de los equipos de protección durante el despeje de las fallas.

La exploración visual se utilizó para estudiar los descriptores de tensión y los descriptores de corriente por separado en las observaciones generadas inicialmente para las fallas monofásicas del sistema. En la etapa de arranque se tomaron las observaciones de falla monofásica con resistencia de falla de 5Ω y se realizaron representaciones en tres y dos dimensiones, mediante combinaciones de los descriptores considerados inicialmente. La figura 6.4 representa una de las distribuciones representadas al utilizar observaciones con los descriptores de corriente pb_I, ps_I, y h_I. La idea con este análisis era establecer la forma y la cantidad de cúmulos representativos, relacionados con la ubicación de las observaciones dentro del sistema. El concepto de cúmulo o *cluster* se define como el área que encierra o aglomera una cantidad determinada de observaciones del sistema dentro del espacio muestral.

Según las representaciones realizadas, los cúmulos están relacionados directamente con el término *zona* del sistema. Una zona comprende la agrupación de un número determinado de nodos dentro del sistema. La presencia de nodos en cada zona depende de su ubicación dentro del sistema estudiado. A través del concepto de zonas, podría establecerse una zona por cada nodo del sistema, lo cual representaría un modelo bastante descriptivo del sistema. Análisis de conglomerados realizados por medio del algoritmo *k-means*, demostraron que la mejor manera de crear zonas dentro de las distribuciones, estaba relacionado con el hecho de agrupar un número mayor de nodos por cada zona establecida. El algoritmo *k-means* exige un volumen mínimo a cada grupo,

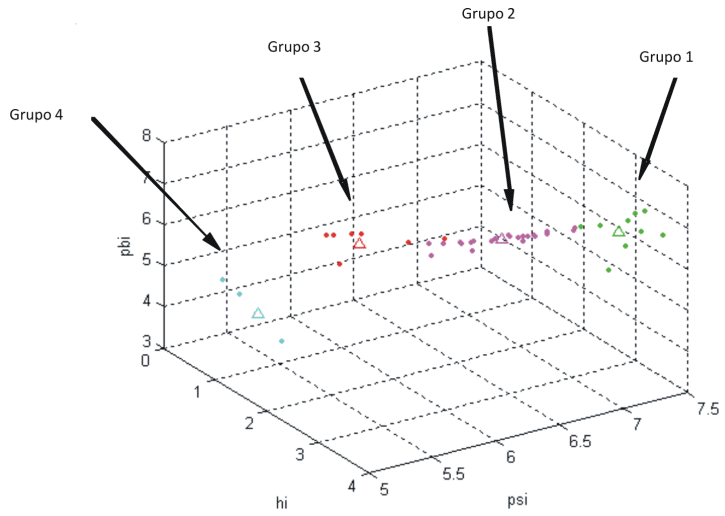

Figura 6.4: Distribución de datos de fallas monofásicas con resistencia de falla de 5Ω

que asegure la no singularidad de las matrices de varianzas y covarianzas calculadas para cada grupo. Esta medida asegura la convergencia del algoritmo al realizar el cálculo de los centros y las matrices de varianzas y covarianzas asignadas a cada grupo. Se puede apreciar la distribución de datos utilizada y una primera agrupación de las observaciones de falla en cuatro grupos principales (ver figura 6.5). Los puntos señalados con rojo representan los centros de los grupos estimados por el algoritmo *k-means*.

A través de exploraciones mayores, se tomaron todas las observaciones de falla monofásica con diferentes valores de resistencia de falla, para observar si se conservaba el mismo patrón encontrado anteriormente. Las exploraciones realizadas mostraron que la forma de la distribución de los datos era similar para cada valor de resistencia de falla. Sin embargo, al aumentar el valor de resistencia de falla, las observaciones correspondientes realizaban un corrimiento en relación con la distribución de resistencia de falla de 5Ω. Este corrimiento ubicaba datos con valores mayores de resistencia de falla, en posiciones similares a observaciones correspondientes a nodos más alejados del punto de alimentación pero con resistencia de falla menor. Al aumentar los valores de resistencia de falla en cualquier punto del sistema, las característi-

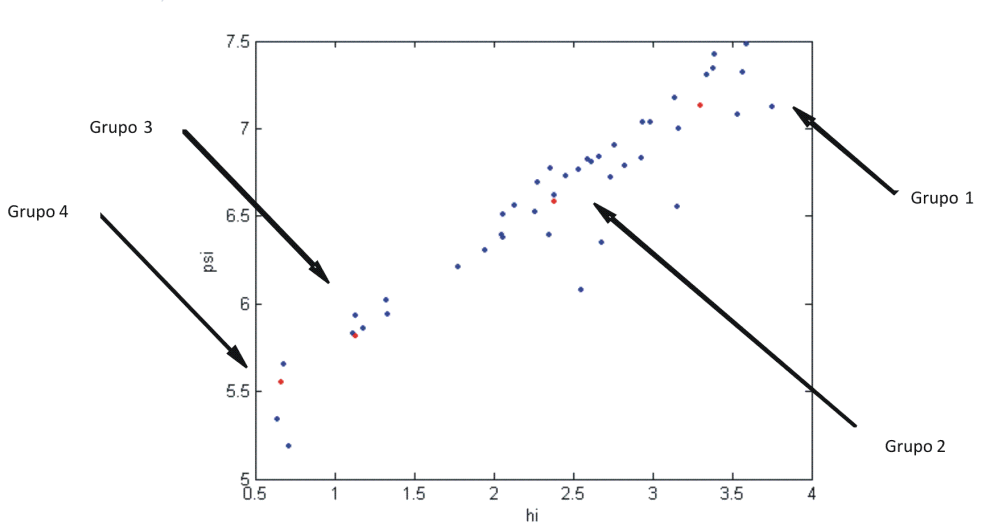

Figura 6.5: Distribución de datos de entrenamiento (puntos azules) y centros de grupos estimados

cas de las señales de falla registradas, decrecen en su magnitud. Por lo tanto el pico de corriente registrado para una falla en uno de los nodos intermedios del sistema con valores altos de resistencia de falla, puede ser similar al pico de corriente registrado para el mismo tipo de falla, pero con valores menores de resistencia de falla, en un nodo bastante alejado de la cabecera el circuito. La figura 6.6 ilustra representación de la distribución de observaciones de fallas con diferentes valores de resistencia de puesta a tierra, generados por los descriptores disponibles. Además se representa las distribuciones separadas según los rangos de cada resistencia de falla para entender el fenómeno del corrimiento.

Pruebas posteriores descartaron el uso de los descriptores relacionados con las pendientes de ascenso y descenso (pb_I, ps_I, pb_V y ps_V) en las señales de tensión y corriente, al punto de utilizar solamente los valores de picos de corriente h_I y huecos de tensión h_V ($sags$) de cada fase en valores por unidad, como observaciones de análisis.

El uso de valores en por unidad permite utilizar los descriptores bajo la misma escala en el espacio vectorial donde se definen los datos multivariantes utiliza-

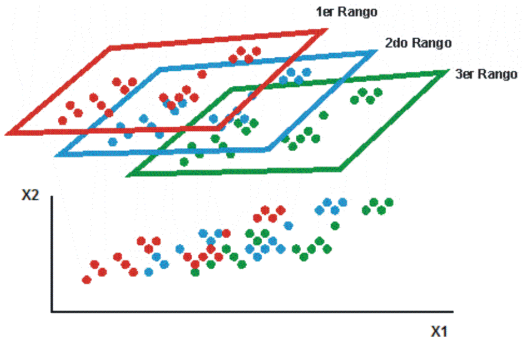

Figura 6.6: Distribución de observaciones de falla con diferentes valores de resistencia de falla

dos, ya que manejamos dos cantidades que normalmente poseen dimensiones diferentes. Los descriptores seleccionados permitieron realizar exploraciones visuales sin el precedente de los corrimientos debidos a los diferentes valores de resistencia de falla.

Las exploraciones iniciales mediante el uso del nuevo paquete de descriptores permitió observar los datos en distribuciones con conglomerados que representaban los diferentes valores de resistencias de falla. En las representaciones de estas distribuciones, aparece una serie de brazos o ramales formando una figura parecida a una mano. En la figura 6.7 se representa la distribución de observaciones de falla monofásica utilizando los descriptores h_V de las tres fases. Cada ramal representa los datos con igual valor de resistencia de falla asociado. Las observaciones de cada ramal corresponden a datos de cada una de las barras del sistema.

Sin embargo en las representaciones realizadas sólo se utiliza una parte de los descriptores disponibles. Las visualizaciones en tres dimensiones limitan el uso de los seis descriptores disponibles para el estudio. Utilizando otras técnicas de análisis gráfico de datos multivariantes, tales como las coordenadas paralelas (ver anexo D), fue posible observar con todos los descriptores las tendencias de las observaciones. A través del comando *csparallel* de MATLAB fue posible realizar las representaciones en coordenadas paralelas de las distribuciones de los datos para los diferentes tipos de falla. En la figura 6.8 se

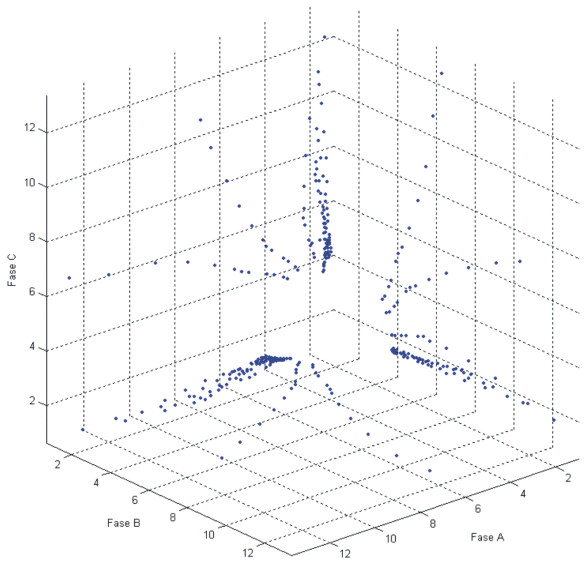

Figura 6.7: Distribución de datos entrenamiento correspondiente con fallas monofásicas del sistema estudiado

presenta las representaciones en coordenadas paralelas de tres conglomerados hallados dentro de las observaciones de falla bifásica, relacionados con tres valores diferentes de resistencia de falla.

6.4.3. Transformación de los datos mediante disminución de dimensiones

Otra forma de utilizar todos los descriptores y utilizar la exploración visual directa, se realizó mediante el uso de transformaciones para disminuir el número de descriptores de las observaciones. Las transformaciones a utilizadas no debían afectar la variabilidad de las observaciones utilizadas. Cuando se manipulan los datos y la variabilidad de los mismos se ve afectada, la correlación que existe entre éstos cambia. Por tanto, un cambio en la correlación de las variables en estudio, altera la información inmersa en los datos, y disfraza los resultados obtenidos al utilizar diferentes métodos de inferencia y estimación.

Primero se realizaron pruebas con operaciones entre los descriptores selec-

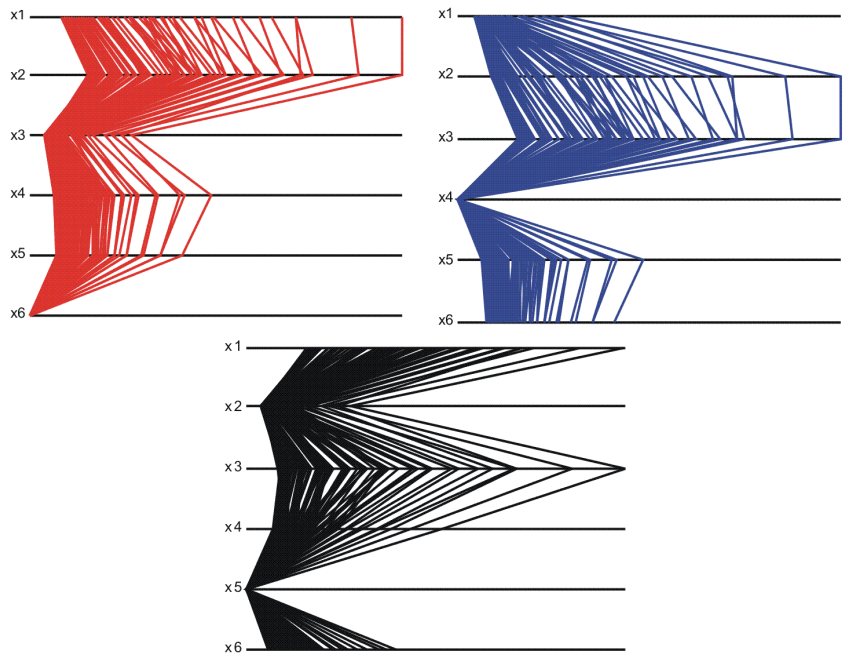

Figura 6.8: Representación en coordenadas paralelas, de observaciones bifásicas doble línea a tierra

cionados para generar distribuciones a partir de las transformaciones utilizadas, sin afectar la variabilidad de éstos. Las pruebas realizadas inicialmente en observaciones de fallas monofásicas permitieron pasar de seis a sólo dos descriptores. El uso de estas transformaciones demostró que a partir de tres estadísticos de tensión y tres estadísticos de corriente (un total de seis descriptores), se utilizarían sólo dos descriptores después de realizar las transformaciones que producen las distribuciones bidimensionales (figura 6.9). Los nuevos descriptores calculados serán útiles durante la etapa de construcción de los modelos y estimación de los posibles lugares de falla dentro del sistema. El cambio de dimensión exige tres transformaciones diferentes, según la fase fallada de la observación.

Un análisis similar utilizando el mismo tipo de transformación para los demás tipos de falla, mostró que las transformaciones utilizadas para las fallas monofásicas, no eran compatibles con el resto. Por lo que debía aplicarse transformaciones diferentes dependiendo del tipo de falla. Sin embargo, cada tipo

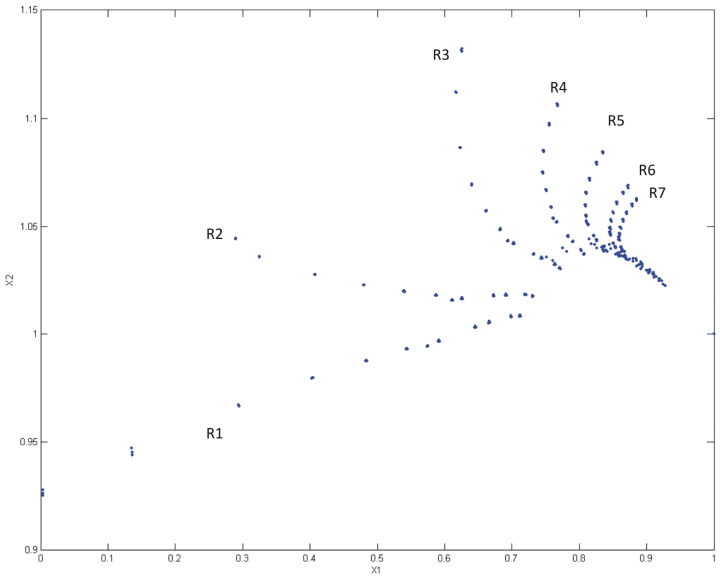

Figura 6.9: Distribución de observaciones de fallas monofásicas según rango de resistencia de falla

de transformación encontrada sólo difiere del orden de uso de los descriptores utilizados. En sí se mantiene el mismo concepto.

Un segundo grupo de pruebas se realizó a través del uso de la técnica de los componentes principales. Mediante la aplicación de los componentes principales fue posible disminuir el número de variables utilizadas en los algoritmos y llevar a un mismo plano las tres distribuciones formadas por observaciones del mismo tipo de falla pero con fases falladas diferentes.

Para utilizar la técnica de componentes principales como herramienta de ayuda, es necesario separar los datos de entrenamiento según las fases falladas y utilizar esta selección para calcular su matriz de varianzas y covarianzas. Los valores propios y los vectores propios de la matriz de varianzas y covarianzas serán de ayuda para transformar los datos en observaciones de una dimensión menor.

A través de la identificación de los valores propios más significativos, seleccio-

namos los vectores propios asociados a estos valores propios que definen las nuevas variables.

Cada matriz de varianzas y covarianzas generará sus vectores propios y valores propios. En total se dispondrán de tres conjuntos de vectores propios, para llevar las diferentes observaciones a un plano en común. Mediantes los seis descriptores utilizados se obtendrá un nuevo paquete de observaciones con sólo dos descriptores los cuales facilitarán la operación de los algoritmos y la visualización de los resultados, sin pérdida significativa de los datos. (ver anexo C).

Los datos de diferentes fases falladas para un mismo tipo de falla son llevados a un mismo plano y su ubicación concuerda con fallas registradas en fases diferentes, en la misma ubicación dentro del sistema estudiado (ver capítulo 7).

La evaluación de las dos técnicas realizadas para disminuir la dimensión de los datos observados mediante análisis de correspondencias, permitió escoger el método de los componentes principales como la técnica más eficiente para realizar las transformaciones.

6.5. Nivel I: clasificación según fase fallada

La primera etapa o nivel comprende la clasificación de los datos según la fase o fases falladas del sistema. El algoritmo *k-means* nos proporciona los centros iniciales necesarios para el inicio de las estimaciones del algoritmo *EM* (*Expectation-Maximitation*). Las pruebas realizadas en la primera etapa mediante modelos *homocedásticos*[3] y *heterocedásticos*[4] al utilizar el algoritmo *EM*, demostraron mejores resultados con modelos heterocedásticos. El hecho de permitir al algoritmo *EM* generar por sí mismo cada uno de los parámetros

[3]Modelos con varianza constante entre componentes. Se caracteriza por formación de grupos esféricos.

[4]Modelos con diferentes varianzas definidas para sus componentes. Generalmente poseen grupos en forma de elipses cuyo eje mayor, se dirige hacia la componente de mayor variabilidad.

de cada *cluster* que definen los grupos de la mezcla finita, libera al analista de establecer los valores de las matrices de covarianzas y los coeficientes de mezclado que mejor se adapten a la distribución analizada. Lo anterior se apoya en dos aspectos: El primero tiene que ver con la forma y cobertura de los grupos generados y el segundo involucra al analista, que utiliza estos métodos. Al utilizar mezclas de distribuciones normales, cada grupo adopta una forma circular u oval. Esta forma depende de la variabilidad de los datos de la distribución, que pertenecen a cada grupo.

Los datos uniformemente distribuidos presentarán una forma muy circular de su distribución, en comparación de aquellas distribuciones que presentan una variabilidad mucho mayor entre sus datos, por lo que la forma de este conglomerado se estirará por el eje de mayor variabilidad y adquiere una forma ovalada representativa.

El uso de modelos heterocedásticos mediante el algoritmo *EM* permite calcular modelos mucho más acordes con las distribuciones de los datos, el problema reside en encontrar buenos puntos de partida de los parámetros necesarios para generarlos. Cuando se poseen distribuciones con conglomerados de clara identificación, resulta sencillo para el analista escoger parámetros de inicio muy simples como matrices identidad para las matrices de covarianzas y coeficientes de mezclado uniformes. Los centros iniciales son generados por el algoritmo *k-means*. El algoritmo *EM* se encarga de comparar los parámetros suministrados, con las distribuciones de los datos bajo análisis. Iteración por iteración va adaptando los parámetros a las características halladas en la distribución de datos, hasta que la solución converge en una serie de parámetros ajustados a los datos suministrados. Cuando las distribuciones de los datos son mucho más difusas y los grupos dentro de la distribución no puedan apreciarse claramente, el analista debe proporcionar parámetros de inicio más precisos para lograr mejores resultados. Lo anterior implica un amplio conocimiento y gran esfuerzo por parte del analista para establecer parámetros lo suficientemente cercanos a los parámetros que en realidad representan y describen adecuadamente la distribución de los datos que está analizando.

El uso de modelos homocedásticos en este tipo de situaciones facilita la labor del analista. Por un lado estos modelos establecen un patrón el cual será uniforme para todos los grupos generados, dicho patrón se generará una sola vez y será copiado a todos los grupos de la distribución. El resultado generalmente son grupos de forma circular que pueden tener o no el mismo tamaño. La diferencia la realiza el algoritmo *k-means*, el cual determina puntos de inicio para los centros de los grupos muy próximos a los posibles centros reales de la distribución y los coeficientes de mezclado escogido de acuerdo a la importancia del grupo dentro de la mezcla. El uso de modelos homocedásticos disminuye la precisión de los modelos generados para describir el comportamiento de los datos bajo análisis, pero presentan una forma muy sencilla de generarlos.

Al observar los datos presentes en la distribución de fallas bifásicas de la figura 6.10, se aprecia claramente tres grupos representativos. Escoger matrices identidad, como matrices iniciales de covarianza para cada uno de los grupos y coeficientes de mezclado de igual valor para estas distribuciones, mostró buenos resultados. Los centros iniciales de los grupos fueron suministrados previamente por el algoritmo *k-means*.

Los grupos generados por el algoritmo *EM* bajo las condiciones descritas, representan apropiadamente cada uno de los conglomerados hallados en la distribución de datos analizada (ver capítulo 7). En la figura 6.11 se representa la clasificación de datos realizada a través de los parámetros calculados por el algoritmo *EM* de acuerdo a la distribución de los datos de la figura 6.10. Los datos fueron clasificados según la máxima probabilidad de pertenencia a cada grupo dentro de la mezcla (ver capítulo 5), de acuerdo a las fases falladas. Cada color representa un grupo representativo de la mezcla calculada por el algoritmo *EM*.

Dentro de las distribuciones principales de cada tipo de falla, esto es en el *nivel I*, aparecen una serie de ramales o brazos. Cada ramal aparece como

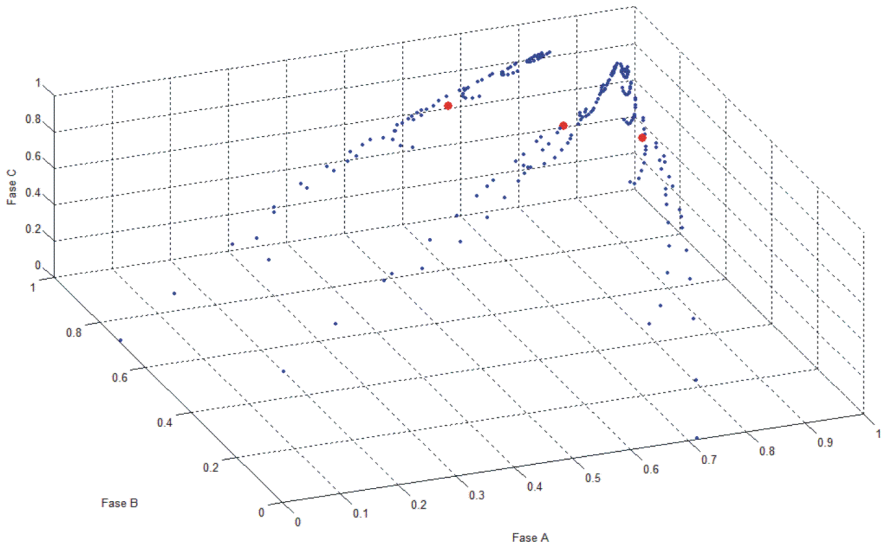

Figura 6.10: *Nivel I* de clasificación: observaciones de fallas bifásicas doble línea a tierra, clasificadas por el algoritmo *EM*

un sub-conglomerado que representa las observaciones generadas con igual valor de resistencia de falla asociado. Finalmente cada observación dentro de estos ramales representa cada uno de los nodos del sistema donde la falla tuvo lugar. Más aún, dentro de cada conglomerado de valores de resistencia de falla también se pueden encontrar conglomerados menores, que agrupan las observaciones de falla. De lo anterior es posible estructurar un segundo análisis de conglomerados, a nivel de resistencias de fallas y un tercer análisis a nivel de observaciones de falla como tal. Estos dos últimos niveles pueden consolidarse como la segunda y tercera etapa en la descripción de los datos dentro de modelos que reúnen la información y los parámetros hallados en cada nivel para efectos de clasificación. A partir de la segunda etapa o *nivel II* de clasificación surgen dos propuestas acerca del orden de análisis y clasificación de los datos para la segunda y tercera etapa.

Propuesta 1

Las distribuciones normales que representan cada zona son establecidas mediante la información previa del rango de resistencia en el cual se clasifica la falla. Por tanto, en la tercera etapa se tiene r mezclas distintas de distribu-

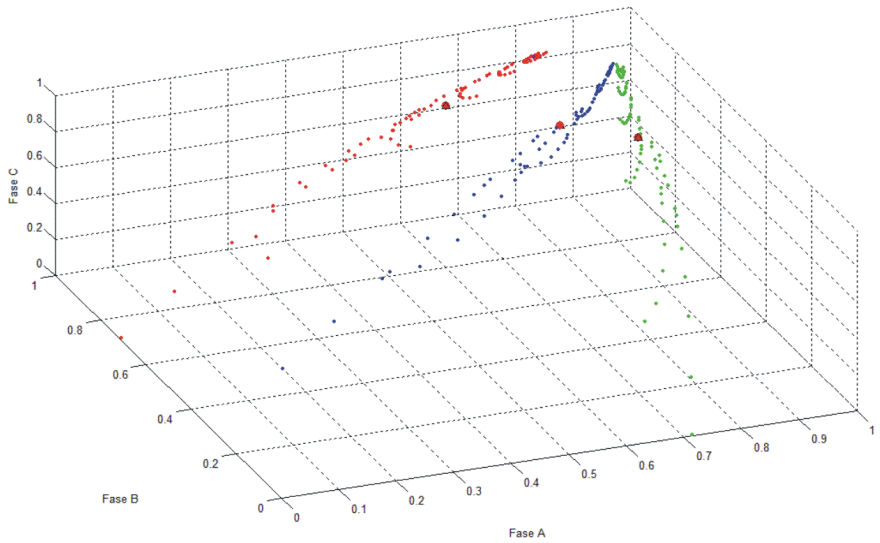

Figura 6.11: *Nivel I* de clasificación: observaciones de fallas bifásicas doble línea a tierra, clasificadas por el algoritmo *EM*

ción normales, cada una derivada de los conglomerados de resistencia de falla previamente determinados. Cada mezcla de distribuciones normales tendrá n grupos definidos, de acuerdo a la selección del número de zonas utilizadas dentro del sistema, tal como se representa en la figura 6.12.

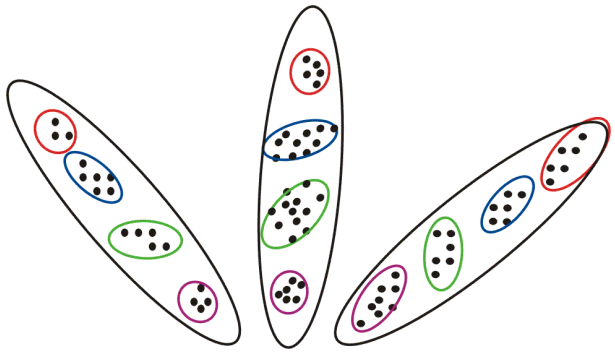

Figura 6.12: Forma de conformar los modelos de mezcla de distribución para localización de fallas, propuesta1

En este tipo de modelo, la discriminación hecha por la etapa de clasificación de resistencias de falla, permite establecer los límites de cada zona para

fallas que ocurran en un rango determinado por la segunda etapa clasificación. Lo anterior permite distinguir entre una falla ocurrida en un punto alejado del alimentador, con un valor alto de resistencia de falla, de una falla cerca del alimentador, con valores bajos de resistencia de falla. El resultado obtenido en el *nivel III* de clasificación dependerá de las estimaciones previas realizadas en la segunda etapa de clasificación. Estos modelos permiten separar cada evento de falla, de otros datos, que aparentemente se ubican en la misma zona, pero que son registrados bajo condiciones diferentes.

Propuesta 2

La propuesta, realiza la clasificación de los datos de manera independiente a las etapas de clasificación anteriores a ella. Este modelo utiliza inmediatamente todos los datos de entrenamiento de cada zona, sin importar el valor asociado de resistencia de falla.

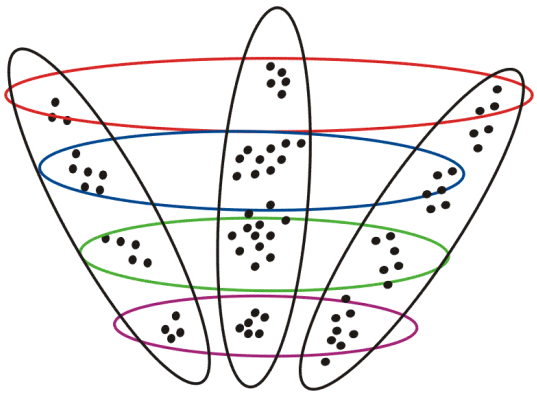

Figura 6.13: Forma de localizar las fallas según propuesta 2

El principal inconveniente de este modelo representado en la figura 6.13, es la pérdida de resolución por parte de las distribuciones en relación con las demás. Lo anterior significa que existen áreas en las cuales la distribución correcta pierde su capacidad de "atrapar" el dato para sí, asignando éste último a la distribución que posee más capacidad de hacerlo. Al momento de determinar los parámetros de la distribución que representa la zona deseada, se pretende describir de la mejor forma todas las características de esta zona.

Sin embargo la expresión "de la mejor manera", no asegura que se alcance a describir toda la zona propuesta, y es aquí en donde aparece el error de asignación. Durante pruebas realizadas para comparar la eficiencia de las dos propuestas, los resultados descartaron la propuesta 2.

6.6. Nivel II: clasificación según resistencia de falla

La información acerca del valor de resistencia de falla permite establecer un posible escenario de falla. Lo anterior hace relación con la capacidad de estimar la clase de objeto que ocasiona la falla en la red. Esta clase de información puede ser relevante a la hora de determinar el tipo de procedimiento a seguir para solucionar el problema.

Una vez identificados los conglomerados que agrupan las observaciones según la fase fallada, se continuó con el análisis de los brazos o ramales resultantes al aplicar las transformaciones en los datos de entrenamiento. Se seleccionó el conjunto de datos correspondientes a la fallas en la fase B, como observaciones de prueba para un análisis de conglomerados y reconocimiento de patrones a nivel de valores de resistencias de falla. El uso del algoritmo *k-means* y el algoritmo *EM* por sí sólos, arrojaron cálculos de grupos diferentes a los esperados. El resultado obtenido fue consecuencia de las definiciones *a priori* dentro de los algoritmos para la formación de los grupos.

Al trabajar con distribuciones normales, se espera que la forma de los grupos a generar tenga una forma oval o circular en el caso más ideal. Al trabajar con las observaciones de prueba, pretendemos formar grupos a partir de una distribución de los datos con altos valores de variabilidad a través de uno de los ejes de distribución de los datos. En consecuencia los algoritmos no estaban programados inicialmente para detectar estas características, y comienzan la búsqueda de cúmulos circulares u ovales de mínima variabilidad según lo preestablecido. La solución a este problema consistió en supervisar parcialmente el análisis de los datos. Mediante la supervisión, el analista interviene de una forma más rigurosa sobre el resultado de las herramientas

estadísticas que utiliza.

En este caso, la supervisión de los resultados consiste en definir vectores de medias y matrices de covarianza para cada grupo, tomando todas las observaciones correspondientes a un valor de resistencia de falla determinado, y sobre este grupo seleccionado, calcular un vector de medias y una matriz de covarianza gracias a los comandos *mean* y *cov*. Estos comandos presentes en MATLAB, permiten calcular los vectores de medias y las matrices de varianzas y covarianza, para el conjunto de datos proporcionados. Mediante el cálculo de estos parámetros, se condiciona la forma de agrupación de los datos dentro de la mezcla de distribuciones.

Inicialmente los coeficientes de mezclado se definen uniformemente para todos los grupos, pues la probabilidad de tener una falla de alta impedancia parece igualmente posible en comparación con fallas de baja impedancia.

Los resultados de las pruebas con los grupos estudiados, demostró que la opción de generar los parámetros de los grupos de forma supervisada, era una buena elección. Sin embargo los grupos generados describían valores de resistencia de falla muy puntuales (figura 6.14). Al realizar pruebas con resistencia de falla de valores intermedios, se presentaba problemas de clasificación. Por ejemplo, si a partir de dos grupos de datos de valores de resistencia de falla de 10Ω y 20Ω respectivamente, se generan dos distribuciones normales (una para cada grupo), el algoritmo de clasificación basado en mezclas finitas, debe clasificar un valor intermedio de resistencia de falla, entre uno de estos dos grupos (el modelo sólo reconoce dos valores puntuales 10Ω y 20Ω). Lo anterior implica que una observación con resistencia de falla de 16Ω, pueda ser clasificada como una falla de 10Ω, si el modelo lo considera así. El paso siguiente fue generar grupos en los cuales las distribuciones de datos se clasificaran por rangos de resistencia de falla. Bajo esta propuesta, un grupo podría describir datos con resistencias de falla asociada con valores entre 10Ω y 20Ω, por ejemplo. Para generar los parámetros de los grupos con rangos de resistencias de falla, sólo se tendría que reunir los datos de entrenamiento

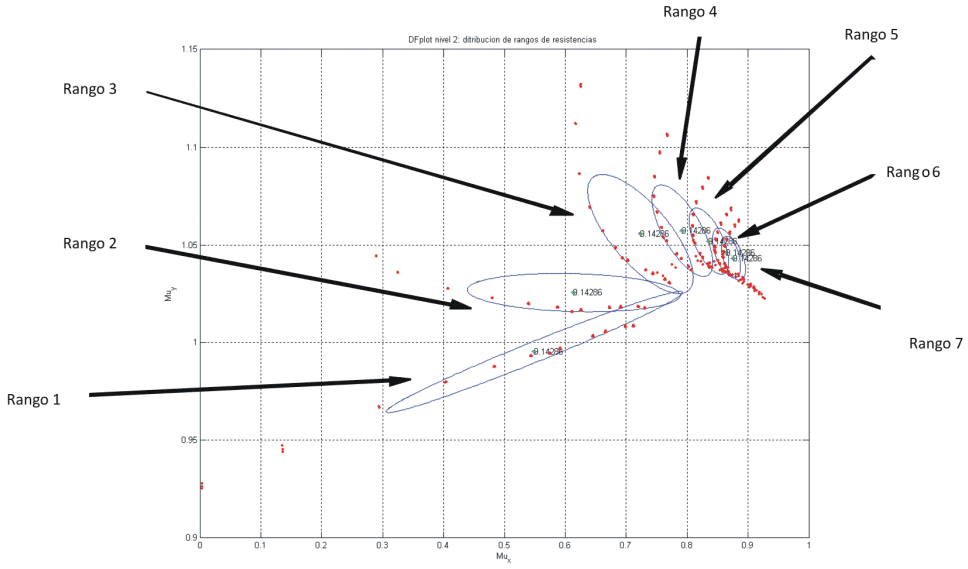

Figura 6.14: Representación DF de7 rangos de resistencias de falla en distribución de observaciones de fallas monofásicas

con las observaciones de falla que contengan lo valores de resistencia de falla dentro del rango determinado; establecer los vectores de medias y las matrices de covarianza para cada grupo de forma supervisada, y definir los coeficientes de mezclado. El resultado de la propuesta, demostró una mayor coherencia en las estimaciones, a pesar que el grado de detalle en la clasificación era menor en comparación con los grupos de valores de resistencias de falla específicos (figura 6.15).

Una vez sometido las observaciones de entrenamiento a las matrices de transformación correspondientes, se inicia nuevamente el proceso de calcular los vectores de media y las matrices de covarianza de los rangos sugeridos de resistencia de falla. En este caso el número de observaciones por cada grupo generado es mucho mayor lo cual mejora la capacidad de cálculo de los parámetros. La disposición de 7 valores diferentes de resistencia de falla, aso-

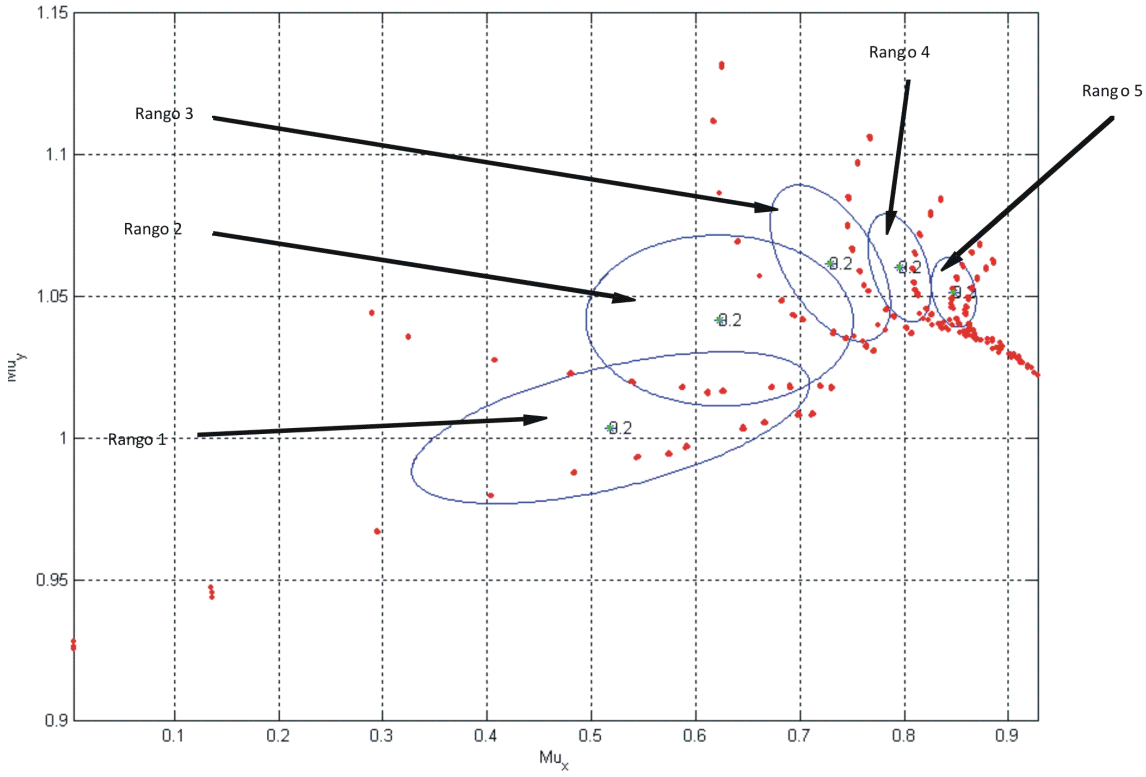

Figura 6.15: Representación DF para 5 rangos de resistencias de falla en distribución de observaciones de fallas monofásicas

ciado a las observaciones de entrenamiento; permite realizar diferentes pruebas para la evaluación del número de rangos a generar. Las pruebas realizadas mostraron mejores resultados con grupos generados por rangos de resistencia de falla entre 10Ω y 20Ω (ver capítulo 7). Generar grupos con rangos muy altos tiene poco sentido a la hora de utilizar la información de la resistencia de falla asociado a un evento registrado. Estimar una falla dentro de un rango de resistencia de falla entre 20Ω y 30Ω es más útil, que estimar la misma falla observada, dentro de un rango de resistencia de falla entre 10Ω y 50Ω. De la misma manera, grupos con rangos muy pequeños de resistencia de falla vuelven muy detallado el modelo, pero surgen problemas en la clasificación eficiente de los datos. Esto se debe a que se crea un número mayor de fronteras entre la distribución de los datos generando un número mayor de zonas críticas de clasificación, para determinar la correspondencia de una observa-

ción dentro de grupos adyacentes.

El problema de clasificar adecuadamente una observación que se halle dentro de los límites de dos grupos siempre estará presente. Disminuir el número de límites generados genera menores inconvenientes de clasificaciones críticas, presentes en dichos límites.

6.7. Nivel III: clasificación según la zona dentro del sistema

El análisis de las observaciones de entrenamiento por etapas permite crear una serie de etiquetas útiles en etapas posteriores. Determinar en cuál fase ocurrió la falla permite etiquetar los datos para conocer qué tipo de transformación aplicarle y así llevar todas las observaciones a un plano en común. Realizar un análisis a nivel de valores de resistencia de falla, discrimina los datos de tal forma que es posible separar todos las observaciones con igual valor de resistencia de falla asociada y observar la información que ofrece para su localización dentro del sistema bajo estudio.

Cada grupo generado dentro del modelo representará un lugar específico del sistema bajo condiciones de falla. Si la base de datos disponible contiene observaciones de falla en cada nodo del sistema, es aceptable asignar un grupo del modelo a cada nodo, aunque nuevamente aparece un análisis muy puntual del fenómeno en estudio. La idea de crear zonas que representen un grupo de barras dentro del sistema permite formar un cuadro más continuo de clasificación. El concepto de zona representado en la figura 6.16, establece un margen en el cual la falla puede ocurrir, no sólo en los nodos pertenecientes a dicha zona, sino a cualquier tramo conductor presente dentro de ésta.

En esta etapa puede utilizarse las distribuciones transformadas en la etapa anterior, pues el interés principal es simplificar el modelo generado. Al mismo tiempo, la creación de etiquetas para cada observación en la etapa de clasificación por resistencia de falla asociada, permite separar los datos y evi-

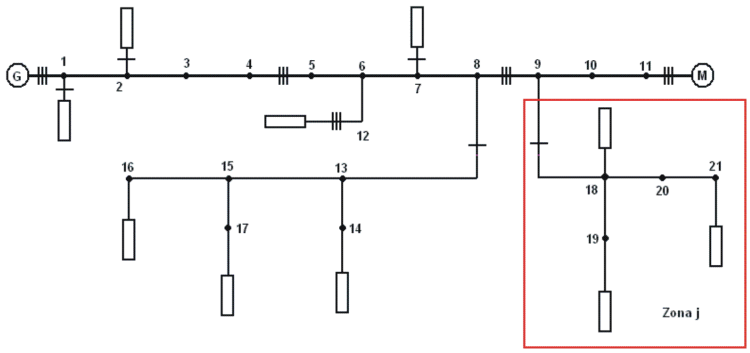

Figura 6.16: Concepto de zona dentro del sistema prototipo

tar confusiones por el corrimiento de los valores de los descriptores. Es claro que debe generarse un modelo de clasificación por zonas, para cada rango por resistencia de falla generado. El uso de nuevas transformaciones en esta etapa, puede alterar la variabilidad de los datos. Generar modelos por cada rango concebido de resistencia de falla, establece una forma mucho más sencilla de calcular los parámetros necesarios para lograr la clasificación de los datos durante esta etapa. La forma de calcular los parámetros de cada uno de los grupos se realiza de manera similar a la etapa anterior. Una vez se conoce cuáles son las observaciones correspondientes a cada rango de resistencia de falla, los datos son separados por su ubicación dentro del sistema y se utilizan para generar los vectores de medias y las matrices de covarianza.

La etapa de clasificación de las observaciones por zonas, comprende el nivel de clasificación más importante. Este es nivel de localización de fallas como tal. Los niveles anteriores sirven para depurar y etiquetar parte de la información presente en la distribución de los datos.

6.8. Selección de modelos utilizando los criterios *BIC*, *ICL* y *AIC*

El empleo de de diferentes niveles para entrenamiento de los algoritmos, permite desglosar poco a poco la información inmersa en las observaciones

disponibles. Para efectos de clasificación, permite situar etiquetas y realizar una descripción de las características de nuevas observaciones de falla obtenidas del sistema estudiado. Los parámetros calculados para cada nivel cumplen con su cometido y realizan discriminaciones acerca de la posible ubicación de nuevos datos suministrados.

Sin embargo el objeto de la aplicación de las mezclas finitas como herramienta de clasificación, consiste en realizar estimados lo más preciso posible. Hallar mejores estimados durante la etapa de clasificación se logra mediante el empleo de mejores modelos clasificadores. Por lo tanto es responsabilidad del analista conformar múltiples modelos y compararlos entre sí para seleccionar aquellos que produzcan los mejores resultados.

Mediante el análisis de múltiples clasificaciones se encontró una relación cercana entre el número de grupos utilizados en los niveles *II* y *III*, con el grado de precisión obtenido en los estimados durante la clasificación de los datos. Mediante el análisis de conglomerados fue posible formar los *clusters* de clasificación para las diferentes etapas. La técnica de análisis de conglomerados no limita absolutamente el número de conglomerados a formar, siempre y cuando no se generen matrices singulares de varianzas y covarianzas. Por lo tanto se necesita un análisis adicional que permita establecer la cantidad óptima de conglomerados. Se puede generar un número excesivo de grupos en busca de una mayor precisión, pero al mismo tiempo, se puede crear información redundante y de poco valor. En el caso contrario, la formación de pocos grupos dentro de la mezcla en busca de un modelo simplificado y de fácil caracterización, podría presentar una descripción muy pobre acerca de las distribuciones, y pasar por alto información importante a la hora de describir los fenómenos estudiados. Además el problema de los límites críticos entre funciones de distribuciones dentro de las mezclas agrega otro elemento que afecta la precisión de la clasificación de los datos.

Los criterios *BIC* (Bayesian Information Criterion), *ICL* (Integrated Clasification Likehood) y *AIC* (Akaike Information Criterium) son herramientas

que permiten evaluar la cantidad permisible de grupos dentro de las mezclas (ver capítulo 5). La formulación de los criterios *BIC*, *ICL* y *AIC*, sirve de referencia para determinar el número óptimo de grupos que describa acertadamente el sistema. Una vez se formulen múltiples modelos con características diferentes, éstos pueden ser evaluados a través de los criterios y comparar los resultados obtenidos. Los modelos cuyos coeficientes calculados por cada criterio obtengan el menor valor, pueden ser etiquetados como aptos para establecer estimaciones lo suficientemente precisas a nivel de clasificación.

La información suministrada por cada criterio es tabulada para determinar los coeficientes de menor valor. Así mismo toda esta información alimenta representaciones gráficas sobre los resultados de los tres criterios de fácil comprensión (figura 6.17). La manera de interpretar estas representaciones consiste en buscar el modelo que posea el menor coeficiente calculado por cada criterio utilizado.

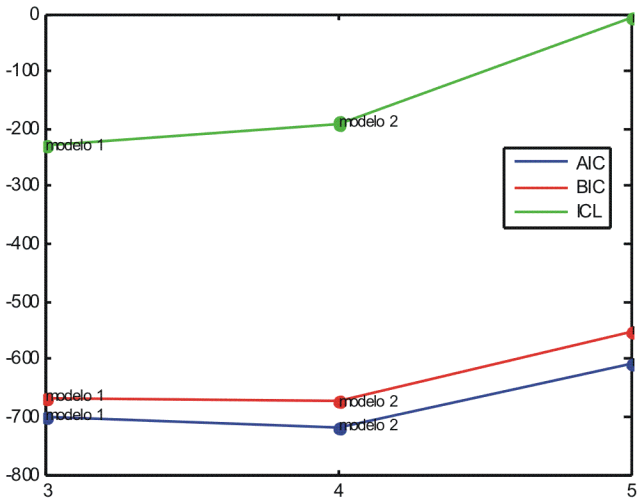

Figura 6.17: Representación gráfica de comparación de modelos aplicando los criterios BIC, ICL y AIC

Con la representación conjunta de los tres criterios, cada línea dibuja la proyección de los coeficientes calculados a medida que el número de *clusters*

aumenta dentro de la mezcla de distribuciones normales. Este tipo de gráficos son útiles para una interpretación rápida de los modelos generados y constituye una etapa anterior al uso de los modelos para la clasificación de nuevos datos disponibles.

6.9. Desarrollo de la propuesta mediante algoritmos implementados en MATLAB

La etapa de análisis, manejo y clasificación de los datos, se desarrolló en paralelo con la construcción de algoritmos en MATLAB que implementaron cada una de las técnicas estadísticas utilizadas. Los programas diseñados a medida que la propuesta se desarrollaba harían parte de un paquete informático capaz de realizar la clasificación estadística mediante mezclas finitas de los datos de falla generados por un sistema de distribución cualquiera.

El software desarrollado debe cumplir con procesar y analizar la información suministrada por el analista para generar modelos que reconozcan los patrones de comportamiento de sistemas de distribución bajo condiciones de falla. Con base en la metodología desarrollada y los algoritmos implementados, el paquete informático debe incluir una serie de funciones que paso a paso culminen en la solución de localización de fallas.

Las funciones planteadas en el desarrollo del software son las siguientes:

- El analista debe suministrar información sobre el tipo de modelos que deben ser generados. Además debe entregar una base de datos cuidadosamente organizada del sistema que pretende estudiar, como observaciones de entrenamiento.
- A partir de datos de entrenamiento, el software debe ser capaz de procesar la información suministrada y generar modelos estadísticos del sistema sometido a análisis de falla. Para ello, se implementaron tareas que contienen las herramientas de análisis estadístico de conglomerados, como la técnica mezclas finitas (MF).

- El software es capaz generar modelos para cada tipo de falla estudiado. Con los tres niveles de clasificación analizados.
- Se implementaron los criterios ICL, BIC y AIC, como ayuda para estimar el mejor modelo a utilizar por parte del analista.
- Realizar la clasificación de nuevas observaciones a través de los modelos generados, para brindar información acerca del posible lugar de falla.
- Generar un reporte donde se almacene la información como producto de la clasificación de los datos ingresados para efectos de evaluación del lugar de falla.

Capítulo 7

Pruebas y resultados

7.1. Introducción

Esta sección presenta la información relacionada con los resultados de las pruebas realizadas durante el desarrollo de la metodología propuesta. Las pruebas comprenden el funcionamiento de los métodos utilizados y la interpretación de los resultados obtenidos utilizando los datos de entrenamiento, así como los resultados producto de la evaluación del conjunto de datos seleccionados para validar la calidad de los modelos clasificadores generados.

El primer enfoque se centra en el análisis de las técnicas gráficas y analíticas de exploración de datos utilizadas para definir las características presentes en los *clusters* formados por las distribuciones de observaciones[1] . En segunda instancia se realizó el estudio de manejo y manipulación de los datos multivariantes. El estudio de la aplicación de múltiples criterios de decisión para selección de modelos, permite entender su forma de operar para convertirlos en una guía eficaz durante la determinación de clasificadores. Los resultados obtenidos se convierten finalmente en la base del desarrollo y aplicación de los conceptos bajo el diseño del paquete informático requerido.

Las pruebas realizadas a las diferentes técnicas se efectuaron mediante el uso de los datos de falla tomados del sistema prototipo utilizado, para efec-

[1]Esta etapa comprende la minería de datos (*Data Mining*) y el reconocimiento de patrones (*Pattern Recognition*)

tos de entrenamiento. El sistema tipo fue dividido en un número de zonas determinado para construir los modelos estadísticos. Se dispone de datos de los diferentes tipos de falla (monofásica, bifásica línea-línea, bifásica a tierra, trifásica, y trifásica a tierra) con diferentes valores de resistencia de falla.

7.2. Análisis de conglomerados

El análisis de conglomerados tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes entre ellos. Inicialmente se tomaron 36 observaciones de fallas monofásicas para evaluar y analizar la capacidad de utilizar los algoritmos *k-means* y *EM* como punto de inicio en el análisis de conglomerados y reconocimiento de patrones.

El análisis de los datos se realizó con tres conjuntos diferentes de observaciones, de acuerdo con los descriptores suministrados a través de la simulación. Se organizaron un conjunto de observaciones con descriptores de tensión únicamente, otro conjunto de descriptores de sólo corriente y un tercer conjunto de descriptores combinados de tensión y corriente.

Antes de iniciar el proceso de análisis y clasificación de los datos se realizó una clasificación manual de los datos seleccionados con el fin de identificarlos dentro de grupos estimados *a priori*, para el conjunto de datos de sólo corriente, sólo tensión, y el conjunto mixto de descriptores de tensión y corriente. Estas distribuciones se contrastaron con los grupos identificados por los algoritmos dentro de las distribuciones de datos utilizados. El objetivo fue entender la forma en la cual el algoritmo *k-means* establece los centros de los grupos, que son utilizados por el algoritmo *EM* para agrupar los datos, en relación con la manera como se organizaron las observaciones a criterio personal.

El proceso inicia con la estimación de las coordenadas de los centros de los grupos dentro de la distribución. En este paso, el algoritmo *k-means* presenta un conjunto de posibles puntos de inicio en los cuales se podrían localizar los centro de los respectivos grupos. La decisión de escoger dichos puntos,

depende del usuario del algoritmo. Se recomienda realizar una representación gráfica donde se puedan comparar los datos de la distribución respecto con la posición de los centros calculados por *k-means*, siempre y cuando dicha representación pueda generarse. En observaciones multivariantes se recomienda el uso de *biplots*[2] , coordenadas paralelas u otras técnicas, para generar gráficos de exploración visual que permitan realizar comparaciones entre todas las variables disponibles.

Si por alguna razón los puntos escogidos por *k-means* no satisfacen las expectativas esperadas, se puede solicitar un nuevo conjunto de puntos centrales hasta obtener los estimados más convenientes. Una representación de los centros estimados contra los datos de falla utilizados, aparece en la figura 7.1. En esta representación los puntos centrales (puntos rojos) se contrastan con las observaciones estudiadas (puntos azules). La representación gráfica corresponde a las observaciones con descriptores de corriente.

Los centros localizados de los grupos mediante el algoritmo *k-means*, alimentan el algoritmo *EM* para realizar el análisis de conglomerados. Los centros finales calculados de los grupos (puntos en verde), y la forma de los grupos dentro de la distribución se pueden observar en las representaciones gráficas DF de las figuras 7.2 y 7.3. En cada gráfica, aparece el valor de las proporciones de los coeficientes de mezcla asignados a cada grupo pg. La forma elíptica de las representaciones de los grupos, está relacionada con las condiciones finales de las matrices de varianza de cada grupo [McLachlan y Peel, 2000]. Las figuras 7.2 y 7.3 representan la misma distribución, desde planos diferentes respecto los descriptores de corriente utilizados en este caso.

El algoritmo *EM* puede configurarse para realizar los cálculos hacia modelos homocedásticos o heterocedásticos según conveniencia del analista. En este caso el algoritmo *EM* tuvo la libertad de manipular los elementos de las matrices de covarianza en cada iteración efectuada (aplicación hacia modelos

[2]Los biplots son representaciones en un plano de dos componentes de observaciones multivariantes. Generalmente se generan múltiples biplots con diferentes combinaciones de los p-componentes, durante el análisis de distribuciones de datos mutivariantes.

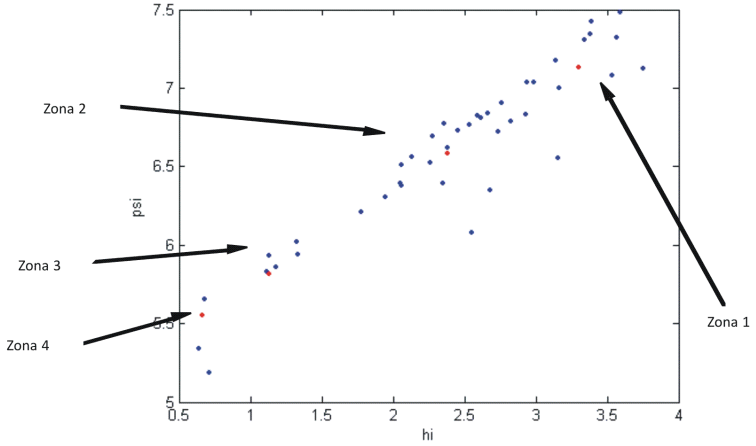

Figura 7.1: Centros iniciales estimados mediante *k-means*

heterocedásticos). Esta libertad, influye en los valores de los coeficientes de correlación entre los descriptores de los datos hasta alcanzar la convergencia predeterminada por el analista. Las matrices de covarianza iniciales de arranque en el algoritmo, se definen como matrices identidad. Si las matrices de covarianzas mantuviera la misma relación entre sí con cada iteración (modelo homocedástico), se obtendría grupos de formas circulares.

Otra manera de observar la forma y tamaño de los grupos es mediante la representación de la función de densidad conjunta de distribución de los datos dentro de la mezcla finita, de acuerdo a la figura 7.4.

Los valores finales para los parámetros de cada grupo aparecen a continuación. Nótese cómo las varianzas de los tres descriptores son muy similares en el primer grupo, por lo que se aprecia una forma casi circular de su representación en las figuras 7.2 y 7.3. En el segundo grupo se observa una mayor varianza con respecto la variable h_I, debido a que el coeficiente de correlación entre h_I y pb_I es mayor respecto a los demás. En la figura 7.3, se observa una forma más plana del grupo en la distribución pb_I vs h_I. En los dos últimos grupos, existe un comportamiento similar respecto al segundo grupo.

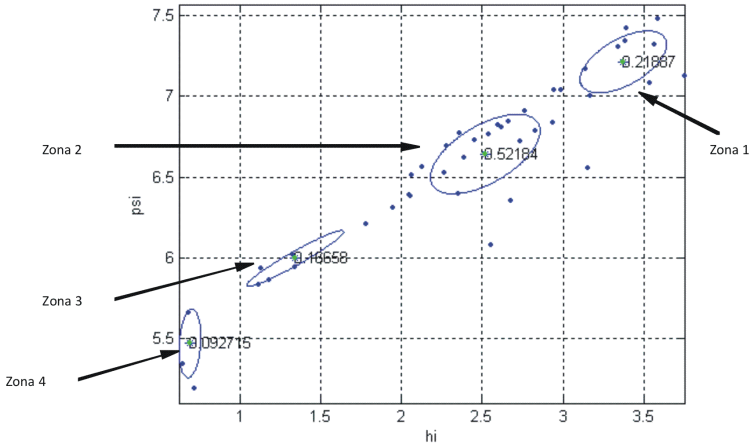

Figura 7.2: Forma de los *clusters* y centros finales. Representación h_I vs ps_I

Matrices de covarianzas calculadas:

$$\mathbf{S}_{g1} = \begin{bmatrix} 0,071733 & 0,029242 & -0,002110 \\ 0,029242 & 0,036935 & 0,026353 \\ -0,002110 & 0,026353 & 0,090680 \end{bmatrix}$$

$$\mathbf{S}_{g2} = \begin{bmatrix} 0,11561 & 0,04952 & 0,072463 \\ 0,04952 & 0,061008 & 0,024992 \\ 0,072463 & 0,024992 & 0,049097 \end{bmatrix}$$

$$\mathbf{S}_{g3} = \begin{bmatrix} 0,090476 & 0,050296 & 0,051297 \\ 0,050296 & 0,030205 & 0,016523 \\ 0,051297 & 0,016523 & 0,013306 \end{bmatrix}$$

$$\mathbf{S}_{g4} = \begin{bmatrix} 0,004517 & 0,002713 & 0,015312 \\ 0,002713 & 0,045336 & -0,158900 \\ 0,015312 & -0,158900 & 0,732050 \end{bmatrix}$$

Centros de los grupos (vectores de medias relacionados):

$$\begin{aligned} \mu_{g1} &= [3,3732 \quad 7,2140 \quad 6,6456] \\ \mu_{g2} &= [2,5179 \quad 6,6424 \quad 6,2633] \\ \mu_{g3} &= [1,3409 \quad 5,9951 \quad 5,7742] \\ \mu_{g4} &= [0,6891 \quad 5,4675 \quad 3,9994] \end{aligned}$$

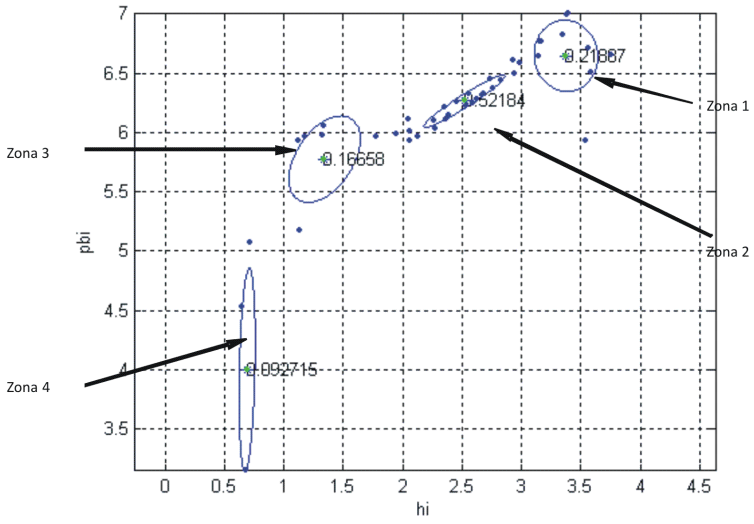

Figura 7.3: Forma de los *clusters* y centros finales. Representación h_I vs pb_I

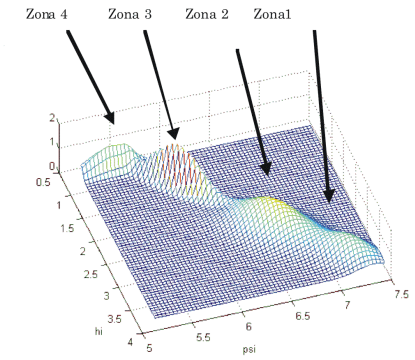

Figura 7.4: Función de densidad de distribución para la muestra analizada

Coeficientes de proporción de mezcla para cada grupo:

$$\pi_g = [0{,}21887 \qquad 0{,}52184 \qquad 0{,}092715 \qquad 0{,}16658]$$

Al final se comparó el resultado obtenido por los algoritmos contra la clasificación de los datos efectuada manualmente. En las representaciones gráficas generadas en las figuras 7.5 y 7.6, cada color representa una zona dentro del sistema de distribución donde fueron agrupados los datos. El valor del color asignado para cada grupo es el siguiente:

- Zona 1: verde.
- Zona 2: magenta.
- Zona 3: rojo.
- Zona 4: cyan.

En la figura 7.5, se observa la clasificación de los datos realizados de forma manual, contra el arreglo establecido mediante el eso combinado del algoritmo *k-means* y el algoritmo *EM* de la figura 7.6. En este caso se utilizaron datos de corrientes de falla con los descriptores h_I, ps_I y pb_I.

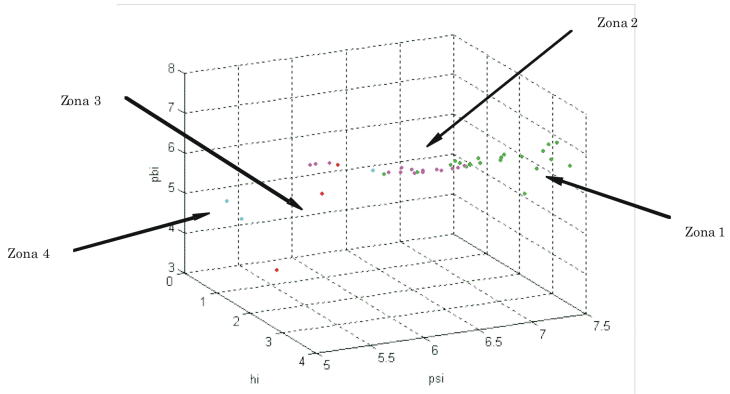

Figura 7.5: Distribución de los datos de entrenamiento, según la clasificación manual previa

La distribución de los *clusters* generados por los algoritmos de análisis de conglomerados presenta una forma más circular en relación con los encontrados

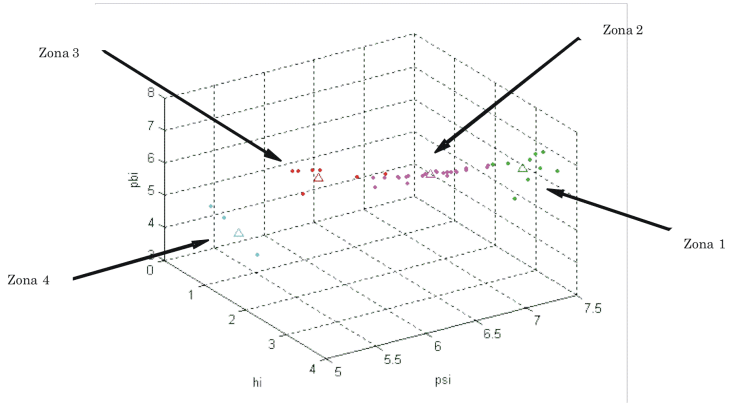

Figura 7.6: Distribución de los datos de entrenamiento, con el uso del algoritmo k-means y el algoritmo *EM*

en las distribuciones de la representación de datos clasificados manualmente. La razón obedece a la búsqueda de grupos con mayor grado de simetría. El grado de similitud alcanzado entre las distribuciones realizadas por los algoritmos en relación con las distribuciones de los grupos realizadas manualmente alcanzó el 84,4 % para los modelos de corriente, 82,6 % para los modelos de tensión y 83,2 % en modelos híbridos de tensión y corriente.

En pruebas de comparación entre la aplicación de modelos homocedásticos contra modelos heterocedásticos, se encontró que el uso de los modelos homocedásticos obedece a la simplificación en las características del modelo generado. En distribuciones cuya exploración visual de los datos refleja conglomerados esféricos, es ideal el empleo de este tipo de modelos. Debido a su rápida convergencia, pueden utilizarse como estimaciones sencillas y para análisis de primera instancia en distribuciones de datos bastante amplias. En el presente análisis, el uso de modelos heterocedásticos es más representativo, pues se desea describir lo más cercano posible la distribución estudiada, aunque implique estimaciones más laboriosas de los modelos requeridos.

En pruebas de eficacia realizadas entre modelos homocedásticos y modelos heterocedásticos en la clasificación de datos de falla, los resultados reflejaron entre un 80 % y 84 % de eficacia en la clasificación de los datos en modelos

heterocedásticos, en relación con eficacias entre el 88 % y 92 % de los modelos heterocedásticos utilizados.

7.2.1. Evaluación como clasificadores de los primeros modelos generados

Una vez establecidos los grupos dentro de la distribución de entrenamiento, el paso siguiente fue verificar la capacidad del modelo generado para localizar nuevas fallas presentes en el sistema de distribución prototipo. Los primeros modelos fueron generados con observaciones de falla monofásicas. Mediante simulación del sistema prototipo en ATP/EMTP se obtuvo los descriptores de cuatro fallas monofásicas en diferentes puntos del sistema para su clasificación mediante mezclas finitas. Los resultados de la clasificación de estas observaciones validaron la capacidad de los clasificadores generados. La resistencia de falla especificada para cada observación es $Z_F = 0,05\Omega$. La información de cada una de las observaciones de falla se describe a continuación:

- Falla1: falla monofásica fase A en el nodo 2.
- Falla2: falla monofásica fase A en el nodo 6.
- Falla3: falla monofásica fase A en el nodo 10.
- Falla4: falla monofásica fase A en el nodo 15.

Los datos utilizados son sometidos a clasificación, mediante la evaluación de la probabilidad de pertenencia de cada observación en los grupos representados por las distribuciones de probabilidad, generadas por el algoritmo *EM* para cada zona (ver expresión 5.45). En la tabla 7.2, aparecen las probabilidades de pertenencia calculadas para cada observación. En la misma tabla se compara la clasificación por zonas de las observaciones a través del uso de las mezclas finitas, contra la ubicación real de las fallas dentro del sistema.

En este caso el modelo con descriptores de corriente fue capaz de clasificar cada observación dentro de las zonas esperadas. Analizando los resultados obtenidos, se puede notar índices muy altos de probabilidad de pertenencia

de los datos cuya localización se encuentra muy cerca de los centros de los grupos a los cuales fueron asignados. Los datos cuya localización se encuentra en el área de confluencia de dos grupos, obtienen índices de probabilidad más bajos, distribuido entre los grupos involucrados. En estos casos el clasificador asignará la observación al grupo que obtenga mayor índice de probabilidad asociado.

En la figura 7.7, se observa la distribución de los datos a través de de la función de densidad conjunta de la mezcla en relación con los descriptores de corriente h_I y ps_I, lo que soporta los resultados obtenidos en la tabla 7.2. Nótese, que las observaciones que se encuentran en la parte cercana de la zona central de los grupos, son aquellos que poseen probabilidades muy altas dentro de dicho grupo. La observación correspondiente a la falla número tres, si bien, fue clasificada dentro de la zona dos como se esperaba, su localización se encuentra alejada del centro del segundo grupo de la mezcla, por lo que el porcentaje de probabilidad de pertenencia se ha distribuido entre los grupos 2 y 3. Sin embargo su posición se encuentra mucho más cerca del segundo grupo, es así que el porcentaje de pertenencia asignado ha sido mayor.

En simulaciones posteriores, se utilizaron un número mayor de datos, aplicados en modelos generados para descriptores de tensión y descriptores de corriente. Los datos de entrenamiento utilizados incluyeron observaciones de falla monofásica en las fases A, B y C. Los clasificadores utilizados para los datos de tensión y para los datos de corriente, dividen el sistema en tres zonas.

Nuevamente los datos de validación utilizados fueron preclasificados de acuerdo con las zonas establecidas, para los grupos especificados dentro del sistema analizado. El modelo basado en descriptores de corriente presentó un comportamiento bastante satisfactorio.

Por otro lado el modelo basado en descriptores de tensión demostró una eficiencia del 84 % en sus estimaciones.

Según los resultados obtenidos parece claro que resulta mejor utilizar modelos basados en solo descriptores de corriente. Sin embargo al utilizar datos de diferente valor de resistencia de falla asociados, aparece una serie de implica-

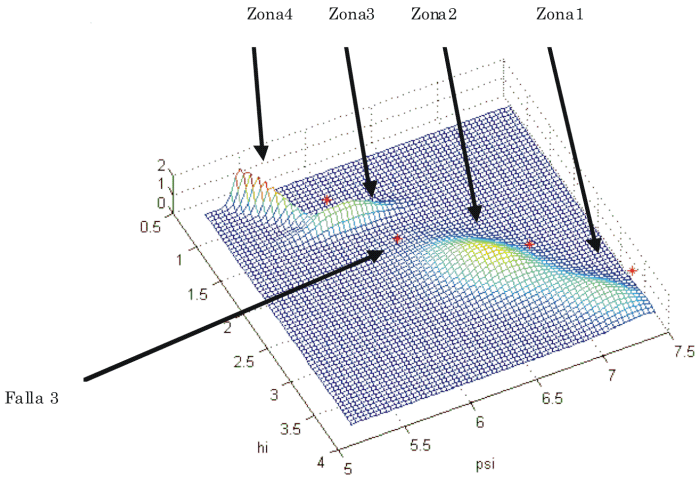

Figura 7.7: Localización de fallas de prueba dentro de la función de densidad conjunta de la mezcla de distribución generada

ciones que vuelve ineficientes los modelos generados. Al utilizar observaciones con diferentes valores de resistencia de falla, la efectividad de los modelos generados decae. Los modelos de corriente producen en promedio un 67 % de eficacia contra 71,3 % de eficacia de los modelos de tensión analizados. Los modelos mixtos presentaron una eficacia intermedia entre los modelos de corriente y los modelos de tensión. En esta instancia los modelos basados en descriptores de tensión poseen resultados positivos en relación con los modelos de corriente estudiados.

Tabla 7.1: Muestra de datos de fallas monofásicas

Observación	$\mathbf{h_I}$	$\mathbf{ps_I}$	$\mathbf{pb_I}$
1	0,576	-14,054	-11.823
2	0,461	-12,334	-10.781
3	0,202	-6,062	-3.180
4	0,093	-3.547	-1.562

Tabla 7.2: Localización de las muestras de falla en el sistema

Falla	Zona Real	Zona Clasificada	Probabilidad por grupo [%]			
			Z_1	Z_2	Z_3	Z_4
1	1	1	99,71	0,18	0,01	0,00
2	2	2	0,30	97,12	2,01	0,59
3	2	2	0,01	56,95	43,02	0,01
4	3	3	0,00	0,00	89,97	10,02

7.3. Transformaciones de los datos

7.3.1. Uso de logaritmos como transformación de los datos

El análisis de un conjunto de datos multivariantes es más simple cuando su distribución es simétrica y las relaciones entre sus variables es lineal. La mayoría de los métodos multivariantes están basados en esta hipótesis. En estas condiciones, la matriz de varianzas y covarianzas es un buen resumen de las relaciones de dependencia existentes.

El logaritmo natural es una de las transformaciones más utilizadas para datos positivos ya que:

- Las distribuciones que describen el tamaño de las cosas, son generalmente muy asimétricas, pero mediante logaritmos la variable puede convertirse aproximadamente en simétrica.
- Cuando las diferencias relativas entre las variables sean importantes, conviene expresar las variables en logaritmos, ya que las diferencias entre logaritmos equivalen a diferencias relativas en la escala original.
- La variabilidad de las variables transformadas es independiente de las unidades de medida.

En la figura 7.9 se representa la distribución de datos de corriente transformados mediante logaritmos, en comparación con la distribución original representada en la figura 7.8.

Tabla 7.3: Descriptores de tensión utilizados para efectos de validación

Falla	**Nodo**	**Fase**	$\mathbf{h_V}$	$\mathbf{pb_V}$	$\mathbf{ps_V}$
1	1	A	0,997	-38,183	58,295
2	3	C	0,709	-33,279	51,407
3	8	A	0,355	-17,443	23,807
4	11	B	0,289	-13,415	18,588
5	15	B	0,177	-19,076	8,241
6	20	C	0,158	-17,4736	17,888

Tabla 7.4: Descriptores de corriente utilizados para efectos de validación

Falla	**Nodo**	**Fase**	$\mathbf{h_I}$	$\mathbf{ps_I}$	$\mathbf{pb_I}$
1	1	A	0,536	13,878	-11,251
2	3	C	0,364	15,619	-10,508
3	8	A	0,175	8,776	-4,760
4	11	B	0,140	6,570	-3,898
5	15	B	0,098	3,284	-3,906
6	20	C	0,089	3,281	-2,229

Tabla 7.5: Clasificación de las observaciones de corriente

Falla	**Zona**	**Probabilidad por grupo** [%]			**Zona**
	Real	$\mathbf{Z_1}$	$\mathbf{Z_2}$	$\mathbf{Z_3}$	**Estimada**
1	1	99,99	0,01	0	1
2	1	99,99	0,01	0	1
3	2	0,089	99,91	0	2
4	2	0,063	99,94	0	2
5	3	0,027	0,61	99,36	3
6	3	0,079	17,04	82,88	3

Tabla 7.6: Clasificación de las observaciones de tensión

Falla	**Zona Real**	**Probabilidad por grupo [%]** $\mathbf{Z_1}$	$\mathbf{Z_2}$	$\mathbf{Z_3}$	**Zona Estimada**
1	1	99,99	0	0	1
2	1	33,84	66,16	0	2
3	2	6,42	85,41	8,16	2
4	2	2,37	88,35	9,28	2
5	3	0,05	0	99,95	3
6	3	99,99	0	0	1

El uso de logaritmos como función de transformación sobre los datos permite obtener una distribución más uniforme, donde los cúmulos de datos se pueden identificar fácilmente. Al utilizarlos directamente sobre los descriptores seleccionados inicialmente ($h_I, h_V, ps_I, ps_V, pb_I$ y pb_V) los cuales poseen dimensiones diferentes, éstos son llevados a una escala relativamente similar, sin perder totalmente las relaciones originales de los datos. El efecto de utilizar los logaritmos le agrega más simetría a las distribuciones pero sin aumentar considerablemente el porcentaje de efectividad de los modelos, sólo se mejora la convergencia de los algoritmos durante la estimación de los parámetros. Sin embargo, estandarizando los valores de cada descriptor en valores por unidad, se logra el mismo resultado de mejoras en la convergencia de los algoritmos, reduciendo en 15 % el número de iteraciones utilizadas para estimación de parámetros.

7.3.2. Transformaciones mediante reducción de dimensiones

Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión r donde $r < p$, es equivalente a sustituir p variables originales por r nuevas variables que resuman óptimamente la información. Esto es posible mediante el uso de componentes principales a través de transformaciones ortogonales [Peña, 2002]. Su utilidad es doble:

- Permite representar óptimamente en un espacio de dimensión menor, observaciones de un espacio dimensionalmente mayor.

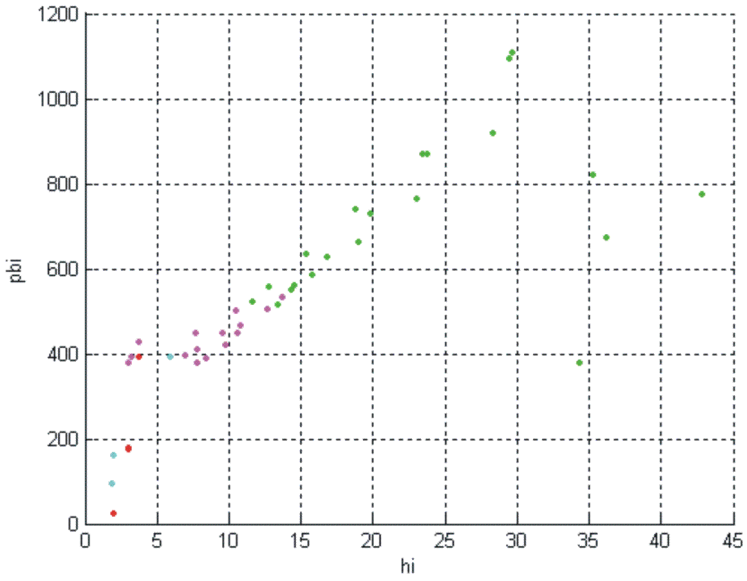

Figura 7.8: Distribución de los datos de entrenamiento, antes de realizar la transformación de éstos

- Permite transformar variables originales, normalmente correlacionadas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

Mediante el uso de las transformaciones se puede llevar toda la información de los datos hacia espacios de dimensiones menores que permitan observar fácilmente las relaciones presentes en las observaciones. Luego de establecer observaciones con tres descriptores de tensión y tres descriptores de corriente, era conveniente disminuir el número de descriptores empleados para el análisis de conglomerados. El empleo de transformaciones para convertir datos de dimensión seis a dimensión dos, mejora la representación gráfica de las observaciones y disminuye el número de parámetros calculados.

Mediante las trasformaciones realizadas por componentes principales fue posible llevar a un plano único las observaciones de diferentes fases falladas para un mismo tipo de falla. Las observaciones de la figura 7.10 representan datos de falla bifásica línea-línea distribuidos en tres conglomerados según las posibles combinaciones de fases falladas, esto es, fallas en fases A y B (distribución

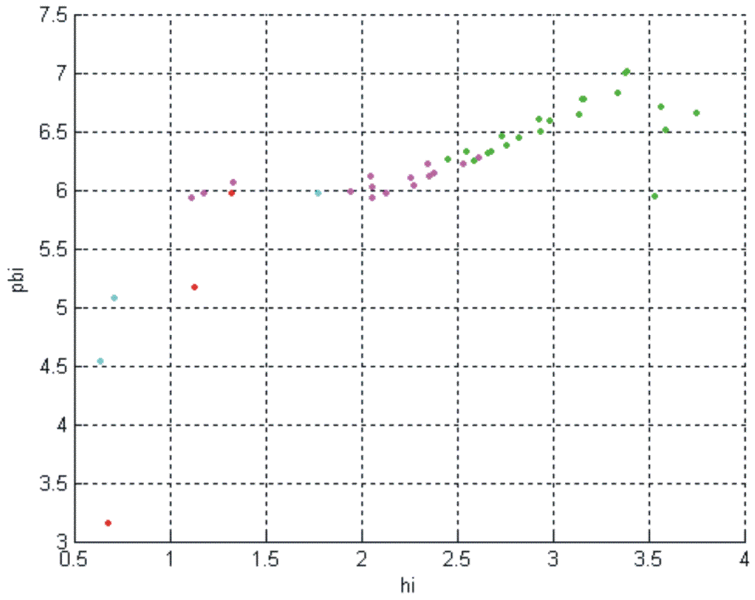

Figura 7.9: Distribución de los datos de entrenamiento, después de realizar la transformación de éstos

azul), fallas en fases B y C (distribución roja); finalmente fallas en fases C y A (distribución verde). Después de aplicar el proceso de transformación por componentes principales, se obtiene la distribución de la figura 7.11. El resultado final es la posibilidad de llevar tres conglomerados de similar distribución con planos de orientación diferentes, a un plano común que muestre la similitud en las características de las observaciones de falla. Las transformaciones efectuadas permiten preparar el conjunto de datos de entrenamiento para los niveles *II* y *III*.

7.4. Revisión de los criterios *BIC*, *ICL* y *AIC*

Los criterios utilizados para evaluar la capacidad de descripción de los modelos a partir de los datos suministrados, parten del uso de la función de verosimilitud descrita en la sección 5. A partir de la función de verosimilitud, cada criterio introduce independientemente, una serie de expresiones que castigan el grado de complejidad que alcance el modelo evaluado.

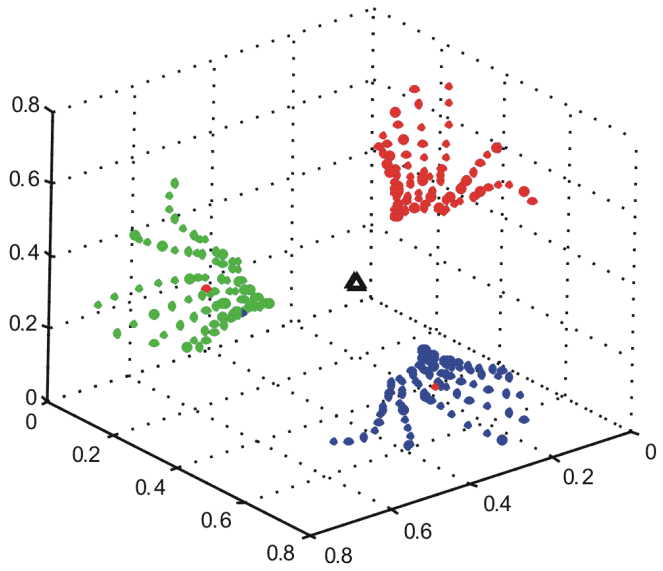

Figura 7.10: Distribución de observaciones de falla bifásica línea-línea según clasificación *nivel I*

Cuando se utilizan las mezclas finitas en análisis de conglomerados, es necesario definir el número de grupos en los cuales se distribuyen las observaciones proporcionadas por la muestra con características similares. En la mayoría de los casos, reconocer a priori la cantidad de grupos representativos de la muestra resulta muy complejo. El procedimiento más común para establecer el número de los grupos, se apoya en la aplicación del criterio *BIC* (Bayesian Information Criterion) [McLachlan y Peel, 2000], como un estimado de la cantidad óptima de grupos presente en el modelo de la mezcla analizada. El procedimiento generalizado consiste en:

- Seleccionar un valor M para el máximo número de grupos.
- Estimar los parámetros para las mezclas con el algoritmo *EM* para $G = 1, 2, ..., M$. En cada prueba, las condiciones iniciales se establecen con un método jerárquico, en este caso el método de *k-means*. Las estimaciones se realizan para todas las posibles condiciones sobre las matrices de covarianzas que se definan.

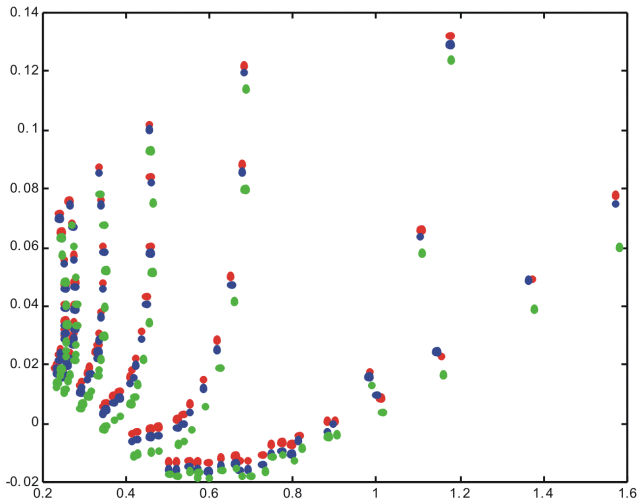

Figura 7.11: Distribución de los datos luego de aplicar transformaciones por reducción de dimensiones en observaciones de falla bifásicas línea-línea

- Seleccionar finalmente el número de grupos y las condiciones de las matrices de covarianzas que minimicen el criterio *BIC*.

Existe una relación entre el número de grupos que considerados y la complejidad de las matrices de covarianza requeridas. Si se permite muchos grupos, se pueden obtener buenos resultados con matrices idénticas del tipo $\sigma^2\mathbf{I}$ (modelos homocedásticos). Por otro lado, con pocos grupos se obtiene mejores resultados con modelos que posean matrices de covarianza diferentes entre sí (modelos heterocedástico).

En la prueba realizada para el criterio *BIC*, se probó un máximo de 10 grupos dentro de una muestra de fallas monofásicas dentro del sistema tipo de 21 nodos y se utilizaron matrices de covarianzas libres para cada grupo. Inicialmente se consideró separar los datos en seis grupos representativos. Cada grupo representa una zona característica del sistema de distribución bajo estudio.
Según los índices calculados con el criterio *BIC* representados en la figura 7.12, el menor valor registrado pertenece al modelo que define 5 grupos para

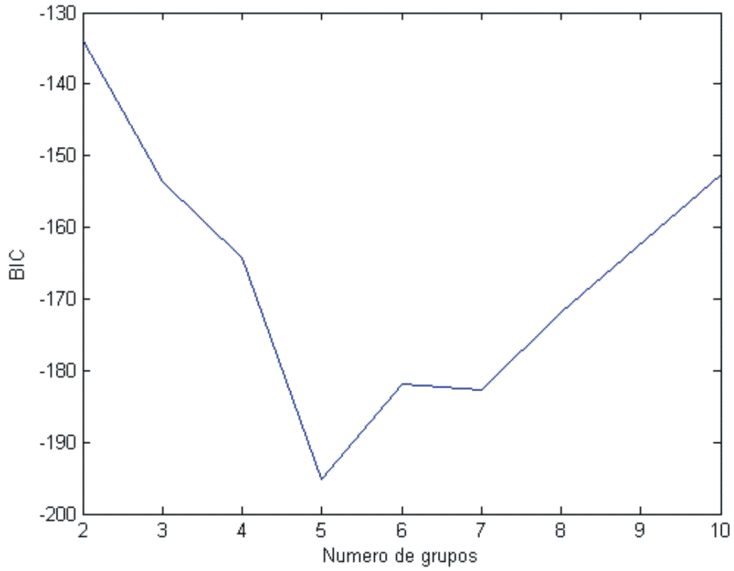

Figura 7.12: Representación gráfica de los índices calculados por el BIC para modelos generados con diferente número de grupos

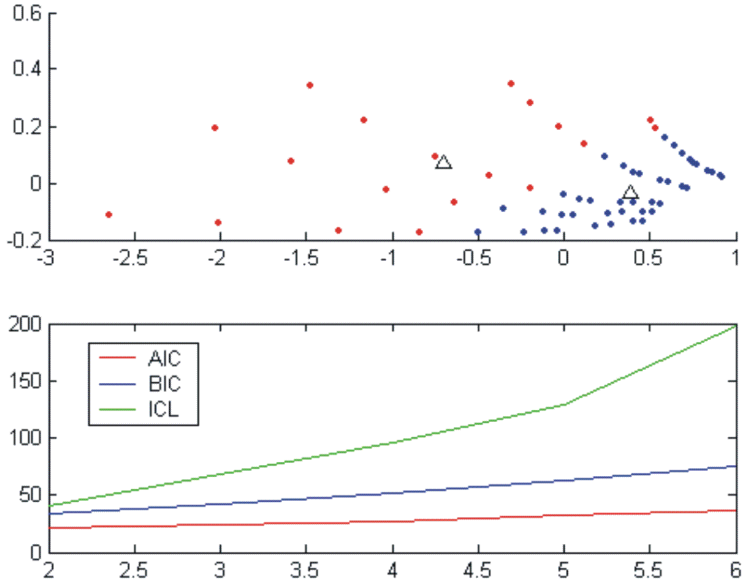

Figura 7.13: Distribución de datos de falla bifásica de acuerdo al modelo sugerido por los criterios BIC, ICL y AIC (arriba). Abajo representación gráfica de los índices calculados por los tres criterios de acuerdo a la tabla 7.7

la distribución analizada. Por otro lado, al acatar los resultados que presenta el empleo del *BIC*, existe la posibilidad de utilizar los modelos con 6 y 7 grupos. Una inspección adicional permitió establecer que la mezcla de distribuciones cambiaría muy poco en su forma a medida que el número de grupos aumenta a partir de $G = 7$. Lo anterior supone que el aumento de grupo no mejora en sí mismo la caracterización del modelo. Las representaciones de las mezclas realizadas muestró una distribución de los datos utilizados, dentro de cuatro zonas principales bien definidas. Retomando la figura 7.12, se aprecia que definir un modelo con cuatro grupos es una buena opción por encima de valores superiores para G. Este resultado se corrobora en la figura 7.1, donde se nota la presencia de las cuatro regiones antes mencionadas. Es posible obtener estimados adicionales aplicando los criterios *ICL* y *AIC* por separado. A menudo se recomienda aplicar más de un criterio a la vez, con el fin de contrastar los resultados obtenidos.

El uso conjunto de los tres criterios puede apreciarse en la evaluación de cinco modelos calculados para descriptores de tensión de fallas bifásicas a tierra registrados en la tabla 7.7

Tabla 7.7: Estimación de índices de concordancia de acuerdo a los criterios BIC, ICL y AIC

No. grupos	**AIC**	**BIC**	**ICL**
2	21,549	34,115	40,603
3	23,869	42,718	68,174
4	27,028	52,16	96,062
5	31,954	63,37	128,868
6	37,221	74,919	198,108

Al generar la representación gráfica de la tabla en conjunto con la distribución de los datos agrupados en los grupos sugeridos, se aprecia la razón de los resultados obtenidos. Debido a la característica de la distribución de los datos, el modelo simple de dos grupos sobresale de los demás modelos comparados.

La simplicidad es una de las características analizada por los criterios. Es-

tablecer una gran cantidad de grupos en un área tan pequeña, implica generar una gran cantidad de parámetros para funciones de distribución de poco alcance. Generar funciones de distribución de mayor alcance y en menor número, permite calcular un número menor de parámetros, representando una disminución de tiempos de cálculo y almacenamiento. Sin embargo existe la posibilidad de distribuciones con más grupos que disminuyen el valor de la función soporte, y son capaces de mitigar la penalidad generada por la gran cantidad de grupos. Por otra parte es posible redefinir los resultados de los criterios si existe la posibilidad de aumentar el número de datos de entrenamiento disponibles que presenten una información adicional de la distribución estudiada. La verificación visual por parte de las representaciones DF y gráficos basados en coordenadas paralelas (ver apéndice D), aparece como una ayuda auxiliar para analizar cada modelo bajo estos casos.

Los datos que aparecen en la tabla 7.8, pertenecen a la evaluación de los criterios *BIC*, *ICL* y *AIC* a dos modelos efectuados utilizando observaciones de falla trifásica con descriptores mixtos de tensión y corriente. La información señala el modelo de tres grupos como el más indicado a utilizar según los tres criterios. Sin embargo los índices calculados reflejan valores muy cercanos y existe la posibilidad de obtener resultados similares con los dos modelos.

Tabla 7.8: Aplicación de criterios BIC, ICL y AIC a dos modelos de eficacia similar para fallas trifásicas

No. grupos	**AIC**	**BIC**	**ICL**
2	-100,797	-96,549	-105,488
3	-104,375	-98,003	-111,678

Una vez se observan los valores tabulados como los presentados en la tabla 7.8, el empleo de los gráficos de distribución DF y las representaciones en coordenadas paralelas permiten en la mayoría de los casos explicar mejor los resultados obtenidos. Analizando el modelo con $G = 3$, es posible observar cómo se distribuyen los centros de los grupos (puntos rojos) a través de la distribución en la figura 7.14. Mediante las representaciones en coordenadas

paralelas, se puede verificar adicionalmente la similitud de los datos clasificados dentro de los grupos conformados por el modelo.

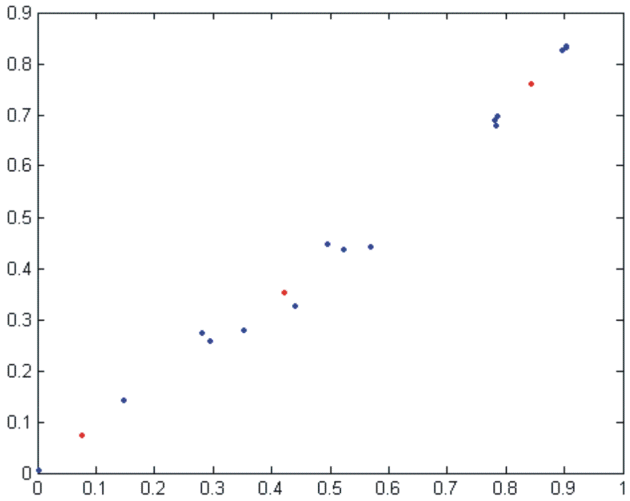

Figura 7.14: Centros (puntos rojos) seleccionados por el algoritmo *k-means* para distribución de observaciones trifásicas

Las representaciones en coordenadas paralelas de la figura 7.15 permite descubrir la tendencia de los datos clasificados dentro de cada grupo y analizar si los grupos conformados reúnen o no las observaciones requeridas. Este análisis puede explicar si los resultados producidos mediante la aplicación de los criterios son consecuentes con lo observado mediante el análisis gráfico.

La necesidad de utilizar más de un criterio obedece a las condiciones establecidas según cada autor, acerca del método para encontrar el modelo óptimo. El contraste entre criterios establece la existencia o no de la uniformidad de los resultados obtenidos. Lo anterior significa que si todos los criterios concuerdan en elegir el mismo modelo, es posible utilizar este modelo para obtener los mejores resultados a diferencia de los demás modelos estudiados. Si por el contrario, existe una divergencia de los criterios en la selección del modelo óptimo, significa que debe realizarse una verificación adicional de acuerdo a las condiciones que cada criterio tiene para establecer sus preferencias.

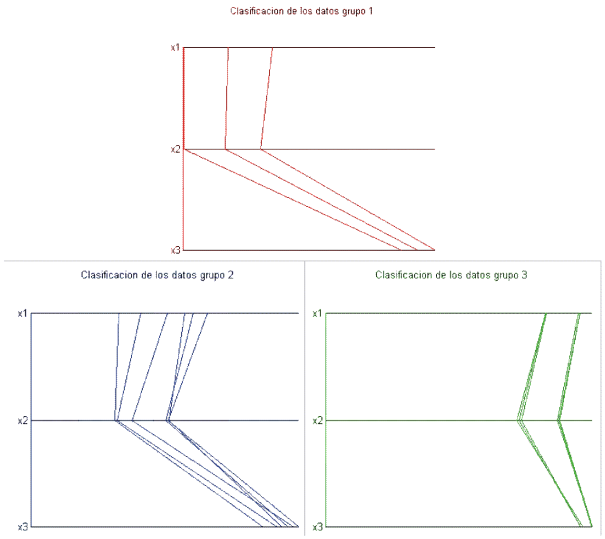

Figura 7.15: Distribución de observaciones en coordenadas paralelas de acuerdo a la clasificación de observaciones por zonas

Las pruebas realizadas confirmaron la preferencia del criterio *AIC* por los modelos con gran cantidad de grupos, rechazando considerablemente los modelos con cantidades pequeñas de grupos. Los criterios *ICL* y *BIC* presentaron resultados bastante uniformes entre sí, en comparación con los resultados del criterio AIC. En el 92 % de los análisis realizados, los criterios *ICL* y *BIC* arrojaron respuestas similares. A primera vista se puede utilizar primero el criterio *AIC* y luego compararlo con el criterio *BIC*. El criterio *ICL* puede utilizarse como una medida para corroborar resultados obtenidos entre los criterios *AIC* y *BIC*. El criterio *AIC* sólo relaciona la función soporte con el número de parámetros utilizados por el modelo estudiado. El criterio *BIC* establece una ponderación del número de parámetros utilizados, a través de la cantidad de datos suministrados para realizar los cálculos con la función soporte. Por otro lado, el criterio *ICL* se comporta como un modelo híbrido entre el criterio *AIC* y el criterio *BIC*. El criterio *ICL* añade el análisis de cada peso π_g utilizado por la mezcla, y lo pondera con el número de observaciones disponibles usadas en la función soporte.

Los criterios *BIC*, *AIC* e *ICL* fueron utilizados para evaluar qué tan cercanos los modelos generados representan las características del sistema en estudio, basados en la función de verosimilitud. Estos criterios sirven como guía para seleccionar el modelo más acotado a los datos utilizados. El uso de los criterios debe acompañarse de la evaluación de los modelos sobresalientes, a través de la cantidad de aciertos que produzcan al estimar la localización de fallas según los datos suministrados para efectos de validación. Para ello, se necesita evaluar su eficacia mediante la medida del porcentaje de observaciones clasificadas correctamente [Martínez *et al*, 2002].

7.5. Pruebas realizadas a los modelos desarrollados mediante el paquete propuesto.

Una vez desarrollado el paquete que utiliza los métodos estudiados para generar modelos capaces de clasificar los datos de falla obtenidos de un sistema de distribución, se evaluó la eficiencia de estos modelos para localizar las fallas mediante los datos de simulación obtenidos del sistema.

Se dispone en total de 1551 observaciones de falla (fallas monofásicas, fallas bifásicas línea a línea, fallas bifásicas a tierra y falla trifásicas). Los datos generados durante la etapa de simulación del sistema bajo condiciones de falla, fueron divididos en dos grupos: datos de entrenamiento y datos de prueba (ver tabla 6.1). Se escogieron 987 datos para efectos de entrenamiento con siete valores diferentes de resistencia de falla asociada. 564 observaciones estuvieron disponibles para efectos de validación, con cuatro valores diferentes de resistencia de falla asociada.

El paquete fue probado con múltiples modelos, variando los rangos de resistencia de falla, el número de rangos, el tamaño de las zonas en el sistema de clasificación y el número de zonas establecidas. El seguimiento del tamaño de las zonas se realizó mediante la numeración de nodos en el sistema estu-

diado.

Se seleccionaron cuatro propuestas diferentes para especificar los rangos de resistencia de falla (*nivel II* de clasificación), y se crearon tres propuestas diferentes para la distribución de zonas (*nivel III* de clasificación) dentro del sistema tipo estudiado. En total se crearon 12 modelos diferentes que combinan las propuestas seleccionadas para los niveles *II* y *III*. El tamaño y forma de las zonas se ilustra en las figuras 7.16, 7.17 y 7.18.
En la etapa de clasificación por rangos de resistencia de falla, se realizaron pruebas con modelos conformados por 3, 4, 5 y 6 rangos diferentes. La forma de distribución de los valores de resistencia de falla para entrenamiento dentro de cada rango, se define en la tabla 7.9. Los valores de resistencia de falla utilizados en la etapa de entrenamiento corresponden a 0,05Ω, 5Ω, 15Ω, 25Ω, 35Ω, 45Ω y 50Ω.

Tabla 7.9: Definición de rangos de resistencia de falla de acuerdo a los modelos propuestos

No. grupos	**Rango 1**	**Rango 2**	**Rango 3**	**Rango 4**	**Rango 5**	**Rango 6**
3	0.05-15	15-35	35-50	-	-	-
4	0.05-15	15-25	25-35	35-50	-	-
5	0.05-5	5-15	15-25	25-35	35-50	-
6	0.05-5	5-15	15-25	25-35	35-45	45-50

A través del paquete *MF_prog*, implementado en MATLAB se calcularon los parámetros para los modelos propuestos (ver anexo A). Estos parámetros fueron almacenados y utilizados posteriormente para realizar la clasificación de las observaciones separadas para efectos de clasificación.

Las reglas de clasificación almacenadas en los modelos, fueron examinadas con los datos de evaluación. Las características de las observaciones de evaluación o clasificación, son desconocidas por el clasificador, por lo tanto, la clasificación de estos datos no será condicionada [Martínez *et al*, 2002]. Una clasificación condicionada aparece cuando se utilizan los datos de entrenamiento para evaluar una correcta clasificación de los modelos generados. En

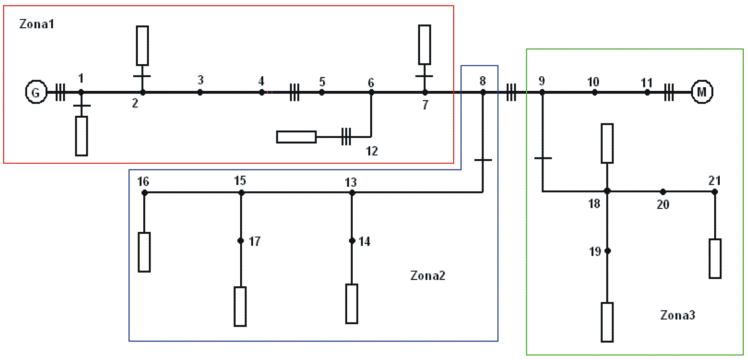

Figura 7.16: Distribución de los nodos del sistema dentro de tres zonas representativas

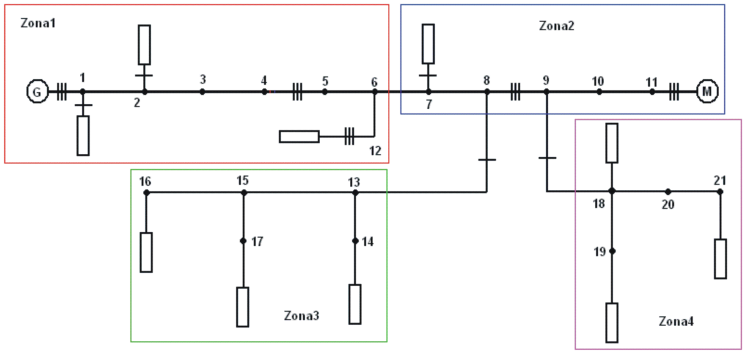

Figura 7.17: Distribución de los nodos del sistema dentro de cuatro zonas representativas

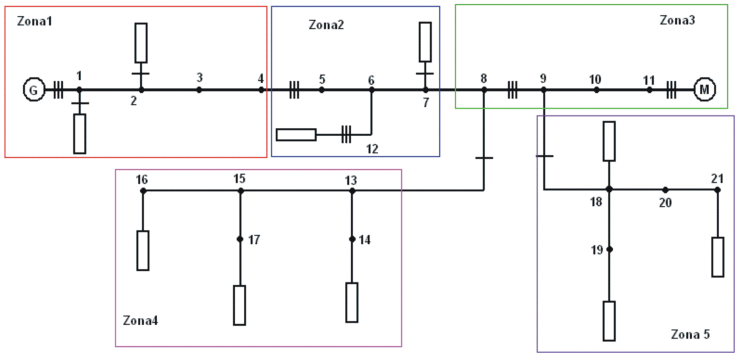

Figura 7.18: Distribución de los nodos del sistema dentro de cinco zonas representativas

este caso, el clasificador ya conoce los patrones de los datos, lo que determina un aumento en la proporción de clasificar correctamente cada observación. Bajo estas condiciones, el clasificador no proveerá una idea exacta de cuál es su capacidad de reconocer patrones en datos que nunca ha visto antes.

Una vez generados los modelos, se utilizaron las observaciones de validación o evaluación para determinar el número de datos correctamente clasificados en cada clase del clasificador. Esta cantidad se denomina N_C. El porcentaje de datos correctamente clasificados se expresa como:

$$p(c) = \frac{N_c}{n_{TEST}} \tag{7.1}$$

Donde n_{TEST} representa el número total de datos disponibles para ser clasificados. Entre mayor sea esta proporción, el clasificador es mejor [Martínez *et al*, 2002].

7.5.1. Verificación del número de grupos dentro de las etapas de clasificación

Una vez calculados y almacenados los parámetros de los modelos propuestos, se utilizaron los criterios *BIC*, *ICL* y *AIC*. Cada criterio utiliza la función de verosimilitud, a manera de verificar cuán cerca se aproximan los modelos al comportamiento de los datos utilizados. La metodología para seleccionar el número de grupos es escoger el modelo con el índice de menor valor calculado por cada criterio [McLachlan y Peel, 2000]. En las tablas 7.10, 7.11, 7.12 y 7.13 se presenta el resultado de los criterios aplicados a los modelos generados para cada tipo de falla. La aplicación de estos criterios se convierte únicamente en una guía y nunca debe tomarse como regla automática de selección [Peña, 2002].

7.5.1.1. Nivel II de clasificación

Si se examinan los resultados de las tablas a nivel general, los tres métodos concuerdan en calcular los índices más bajos para los modelos que utilizan

Tabla 7.10: Aplicación de criterios BIC, ICL y AIC a modelos propuestos para fallas monofásicas nivel II

No. grupos	AIC	BIC	ICL
3	649,67	615,89	862,63
4	651,43	606,40	1165,70
5	657,97	601,68	1522,90
6	648,20	580,66	1961,50

Tabla 7.11: Aplicación de criterios BIC, ICL y AIC a modelos propuestos para fallas bifásicas línea-línea nivel II

No. grupos	AIC	BIC	ICL
3	-377,12	-414,11	-120,47
4	-361,30	-386,41	48,34
5	-394,61	-393,47	346,30
6	-350,10	-412,87	620,74

entre 3 y 4 grupos correspondientes a rangos de resistencia de falla. El criterio ICL, selecciona los modelos con tres grupos por encima de los demás (estos modelos son los más simples). El resultado obtenido concuerda con la definición derivada de la expresión utilizada para este criterio (ver capítulo 5). El criterio *BIC* concuerda con el criterio *ICL* en escoger un número reducido de grupos. Según la expresión que define el criterio *BIC*, al comparar varios modelos con la misma cantidad de datos, aquellos que utilicen más parámetros serán gravemente penalizados. Por otro lado el criterio *AIC* establece diferencias con los otros criterios en los modelos de fallas monofásicas. Según los índices calculados por *AIC* el modelo con cinco grupos de rangos de resistencia de falla debe ser seleccionado como clasificador. La naturaleza misma del criterio AIC tiende a menudo a escoger modelos con gran cantidad de grupos en muestras medianas.

Tabla 7.12: Aplicación de criterios BIC, ICL y AIC a modelos propuestos para fallas bifásicas doble línea a tierra nivel II

No. grupos	AIC	BIC	ICL
3	498,70	466,78	908,52
4	587,12	546,33	1024,11
5	789,23	756,17	2045,78
6	824,33	812,56	2247,08

Tabla 7.13: Aplicación de criterios BIC, ICL y AIC a modelos propuestos para fallas trifásicas nivel II

No. grupos	AIC	BIC	ICL
3	-427,34	-446,25	-99.31
4	-410,12	-466,76	134,88
5	-418,54	-409,96	264,08
6	-318,34	-365,87	313,66

7.5.1.2. Nivel III de clasificación

Los índices obtenidos en la aplicación de los criterios para la evaluación de los modelos según las zonas de agrupación, presentan resultados menos uniformes a los presentados durante la evaluación de rangos de resistencia de falla. En las tablas 7.14, 7.15, 7.16 y 7.17 se presenta el resultado de los criterios aplicados a los modelos generados para cada tipo de falla Sin embargo se mantiene la decisión de utilizar tres grupos dentro de los modelos aplicados a los diferentes tipos de falla.

Tabla 7.14: Aplicación de criterios BIC, ICL y AIC a modelos propuestos para fallas monofásicas nivel III

No. grupos	AIC	BIC	ICL
3	-844,13	-821,33	-131,12
4	-1024,22	-987,82	-155,45
5	-903,95	-912,84	33,74

Tabla 7.15: Aplicación de criterios BIC, ICL y AIC a modelos propuestos para fallas bifásicas línea-línea nivel III

No. grupos	AIC	BIC	ICL
3	44,65	72,14	110,02
2	-61,84	-47,27	-13,23
3	219,04	177,08	239,45

Tabla 7.16: Aplicación de criterios BIC, ICL y AIC a modelos propuestos para fallas bifásicas doble línea a tierra nivel III

No. grupos	AIC	BIC	ICL
3	-46,22	-87,64	41,19
2	-147,18	-156,05	-119,44
3	-86,09	-128,34	51,07

Tabla 7.17: Aplicación de criterios BIC, ICL y AIC a modelos propuestos para fallas trifásicas nivel III

No. grupos	AIC	BIC	ICL
3	-496,56	-532,15	-341,08
2	-548,01	-571,33	-500,43
3	-86,09	-492,21	-241,87

Utilizando los tres criterios respecto a las zonas escogidas como los grupos de las distribuciones de mezcla del *nivel III*, el resultado es similar al encontrado en el análisis de rangos de resistencias. Cada criterio selecciona o escoge los modelos con menor número de grupos, atendiendo la necesidad de manejar distribuciones más sencillas.

En pruebas paralelas, los valores para seis, siete y ocho grupos dentro de los modelos no son lo suficientemente cercanos a los valores establecidos inicialmente en las tablas según cada uno de los criterios.

Según los resultados obtenidos por los tres criterios, los modelos que ofrecen mejores estimaciones en las clasificaciones serán aquellos que manejan distribuciones de zonas de acuerdo con las propuestas de las figuras 7.16, y 7.17.

7.5.2. Clasificación de los datos de validación

Expuestos los modelos a la verificación mediante los criterios *ICL*, *AIC* y *BIC*, el siguiente paso es comprobar la efectividad de clasificación de nuevos datos. Las 564 observaciones separadas inicialmente bajo la etiqueta de datos de validación, serán utilizadas finalmente para estudiar la eficacia de los modelos generados mediante mezclas finitas.

El primer nivel que comprende la identificación de las fases falladas tuvo resultados positivos. Para cada tipo de falla, los clasificadores acertaron alrededor de 99,88 % del total de los datos dispuestos para clasificación, identificando plenamente las fases involucradas en la falla. Las condiciones iniciales suministradas en el empleo del algoritmo *EM*, fueron influyentes en la selección de los parámetros de las distribuciones del primer nivel. El tipo de falla con mayor índice de error durante la evaluación de los datos en *nivel I* de clasificación correspondió a las fallas bifásicas doble línea a tierra. El porcentaje de error alcanzado por los modelos en este tipo de fallas alcanzó el 2 %.

7.5.2.1. Clasificación según la forma de agrupación de rangos de resistencia de falla

Fallas monofásicas

El porcentaje de acierto presentado, se refiere a la cantidad de clasificaciones correctas estimadas, utilizando datos con resistencia de falla asociada con valores de 10Ω, 20Ω, 30Ω y 40Ω respectivamente. La efectividad de cada modelo al clasificar los datos descritos se relaciona en la tabla 7.18. En total se utilizaron 180 observaciones destinadas a validar las estimaciones de cada propuesta.

Tabla 7.18: Porcentaje de acierto de modelos propuestos para fallas monofásicas nivel II

Propuesta	**No. Rangos**	**Porcentaje por grupo** [%]			
		10Ω	**20Ω**	**30Ω**	**40Ω**
1	3	80	80	80	100
2	4	60	80	80	100
3	5	80	80	80	100
4	6	80	80	80	80

En general el porcentaje de aciertos supera el 80 %. Sin embargo se refleja, errores en la clasificación de observaciones con valores bajos de resistencia de falla en contraste con la correcta clasificación, casi total de las observaciones relacionadas con resistencia de falla de 40Ω.

La baja efectividad de de la propuesta 2, se refleja en la definición de rangos más pequeños para valores menores de resistencia de falla y rangos más amplios para valores mayores de resistencia de falla. Una situación similar se presenta en la propuesta 4 con seis rangos de resistencia de falla. En esta propuesta, al incluir un número mayor de grupos con rangos más pequeños de resistencias de falla, el número de fronteras donde los rangos se traslapan entre sí aumenta (figura 7.19). Si una observación se ubica en estas zonas de confluencia entre dos grupos, el clasificador asignará esta observación al grupo que obtenga mayor probabilidad de pertenencia. Sin embargo dependiendo de los parámetros de cada grupo, habrá lugares donde uno de los grupos domine un sector del espacio de congruencia y reclame para sí las observaciones que se ubique en este espacio. Lo anterior implica la posibilidad de asignaciones equivocadas de datos en grupos dominantes, cuando las observaciones pertenecen realmente a los grupos con menor influencia en el espacio de localización de los datos clasificados.

Los resultados de las clasificaciones en contraste con los datos arrojados por los criterios *BIC*, *ICL* y *AIC*, muestran la propuesta 1 como la mejor alternativa. Aunque se obtuvieron resultados similares en porcentajes de clasificación,

la posibilidad de utilizar un modelo más sencillo y con menos parámetros, facilita y hace más rápida la clasificación de los datos. Por otro lado si es necesario realizar una descripción más detallada de los valores registrados de resistencia de falla, la propuesta 3 ofrece una opción intermedia entre el modelo 1 (más sencillo pero poco descriptivo) y el modelo 4 (más descriptivo, pero menos eficiente).

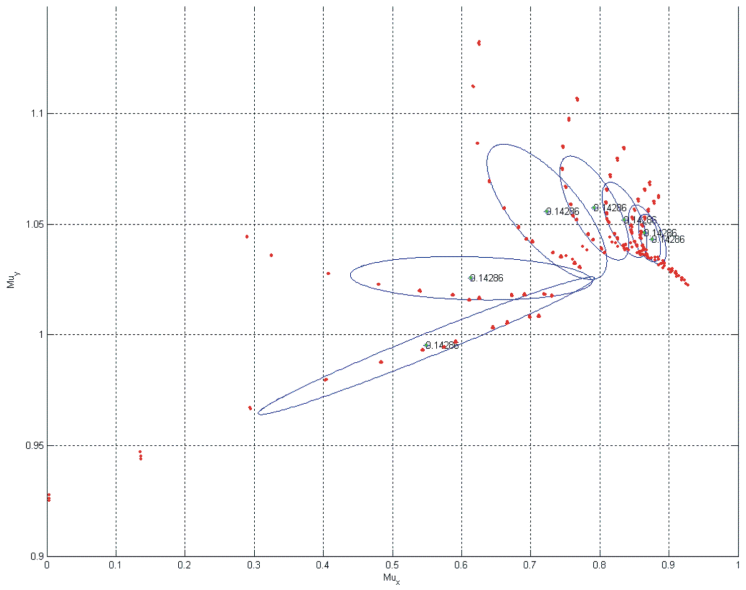

Figura 7.19: Distribución de rangos de resistencia de falla en representación DF

Fallas bifásicas Línea - Línea

El total de observaciones de fallas bifásicas línea-línea utilizadas en la validación de las propuestas fueron 144. Según los porcentajes de acierto descritos en la tabla 7.19, la propuesta 3 refleja los mejores resultados alcanzados. La propuesta 3 para este tipo de fallas, establece un equilibrio entre el tamaño del rango y el total de rangos dispuestos para etiquetar las observaciones clasificadas.

A pesar que el análisis realizado por los criterios *AIC*, *ICL* y *BIC* selecciona las propuestas 1 y 4 como los modelos más representativos, la propuesta 3 alcanza mejores resultados de clasificación. Lo anterior corrobora el hecho de que los criterios son meras guías de rápida revisión, y nunca deben utilizarse

Tabla 7.19: Porcentaje de acierto de modelos propuestos para fallas bifásicas línea-línea nivel II

Propuesta	No. Rangos	Porcentaje por grupo [%] 10Ω	20Ω	30Ω	40Ω
1	3	83,33	100	84,2	92,14
2	4	75	88,88	100	100
3	5	83,33	83,33	100	100
4	6	83,33	83,33	83,33	94,44

como reglas absolutas de selección.

Fallas bifásicas doble línea a tierra

Al igual que las pruebas para falla bifásica línea-línea, se utilizaron 144 observaciones para validar las estimaciones producidas por los modelos generados para fallas bifásicas doble línea a tierra. La tabla 7.20 refleja los porcentajes de acierto alcanzados por cada propuesta.

Tabla 7.20: Porcentaje de acierto de modelos propuestos para fallas bifásicas doble línea a tierra nivel II

Propuesta	No. Rangos	Porcentaje por grupo [%] 10Ω	20Ω	30Ω	40Ω
1	3	80,55	100	86,11	100
2	4	69,44	83,33	100	100
3	5	58,33	77,22	100	100
4	6	58,33	71,66	88,88	86,11

Los resultados muestran un bajo nivel de acierto en las estimaciones realizadas para observaciones con resistencia de falla igual a 10Ω. La propuesta 1 presenta los mejores resultados de clasificación en todos los rangos definidos. Según los porcentajes de acierto, los modelos de las propuestas pierden efectividad a medida que aumentan el número de rangos. Sin embargo, se observa en las propuestas 2 y 3 que las estimaciones realizadas clasifican adecuadamente la totalidad de las observaciones con valores altos de resistencia de

falla asociada utilizados en la validación. Los resultados de las estimaciones realizadas por los criterios *BIC*, *ICL* y *AIC*, corresponden con la información obtenida por la validación de los datos al señalar a la propuesta 1 como la más adecuada para utilizar en el *nivel II* de clasificación.

Fallas trifásicas

Los resultados de las clasificaciones de las observaciones de fallas trifásicas, fueron satisfactorios. Los modelos utilizados de las propuestas fueron lo suficientemente precisos en la selección correcta de las observaciones dentro de los rangos especificados por cada modelo. En la tabla 7.21 se presenta un consolidado del porcentaje de aciertos alcanzados durante la evaluación de 96 observaciones de validación, divididas de acuerdo a los cuatro valores elegidos de resistencia de falla asociada.

Tabla 7.21: Porcentaje de acierto de modelos propuestos para fallas trifásicas nivel II

Propuesta	**No. Rangos**	**Porcentaje por grupo** [%]			
		10Ω	**20Ω**	**30Ω**	**40Ω**
1	3	100	100	100	100
2	4	83	100	100	100
3	5	83	100	100	100
4	6	83	100	100	100

Nuevamente la propuesta 1 obtuvo las mejores estimaciones para los diversos valore de resistencia de falla. En general el problema de estimación se concentró en observaciones con valores bajos de resistencia de falla, en las propuestas que especificaban rangos más pequeños dentro del espacio de valores de resistencia de falla menores.

Los porcentajes de acierto de las propuestas para la evaluación de clasificación para el nivel *II* y nivel *III* bastante similar. Aunque los mejores resultados fueron obtenidos a través del modelo con tres rangos y tres zonas. Este modelo explica muy globalmente al sistema estudiado. Por lo que el uso del modelo con seis rangos y tres zonas en el caso de las fallas trifásicas, presentaría

resultados aceptables con un nivel de precisión un poco menor.

7.5.2.2. Clasificación según la forma de agrupación de nodos en zonas

Para verificar la eficacia de los modelos en este nivel, se observó el comportamiento de los clasificadores por cada tipo de falla. La razón radica en la disposición del circuito estudiado, pues existen tramos netamente monofásicos en los cuales una clasificación de fallas bifásicas o trifásicas no aplica.

Tomando los reportes generados por el paquete *MF_prog* de las evaluaciones que cada modelo efectuó sobre los datos de validación, se alimentaron una serie de tablas que presentan los índices de eficacia de cada clasificador para asignar correctamente las observaciones a la zona que pertenecen.

Fallas monofásicas

En el sistema estudiado, existe la probabilidad de ocurrencia de falla monofásica en cualquier nodo del sistema, por lo tanto se puede definir hasta cinco zonas de agrupación de nodos según las propuestas de las figuras 7.16, 7.17 y 7.18. La tabla 7.22 presenta los porcentajes de acierto alcanzados en cada zona definida por las propuestas, durante la clasificación de las observaciones de validación.

Tabla 7.22: Porcentaje de acierto de modelos propuestos para fallas monofásicas nivel III

Propuesta	**No.**	**Porcentaje por grupo** [%]				
	Zonas	**Zona1**	**Zona2**	**Zona3**	**Zona4**	**Zona5**
1	3	94,60 %	87,78 %	74,33 %	-	-
2	4	95 %	90 %	82 %	79 %	-
3	5	98,88 %	94,50 %	93,67 %	80 %	78,88 %

Al revisar los resultados de los clasificadores respecto a las fallas monofásicas, existe un 96 % de aciertos en la clasificación de observaciones de falla con

valores altos de resistencia de falla (40Ω), contra 87 % de aciertos en clasificaciones con valores bajos de resistencia de falla asociados.

Los mejores resultados fueron establecidos con los modelos 1 y 3. Sin embargo las diferencias cuantitativamente hablando, entre los modelos 2 y 3 no es tan significativa. Bajo otro punto de vista, el modelo 3 se puede interpretar como una alternativa más detallada del modelo 2. El modelo 1 generaliza un poco más la distribución de nodos en las zonas con el fin de obtener mejores resultados en las clasificaciones, bajo apreciaciones más sencillas del sistema.

Fallas bifásicas Línea - Línea

En este tipo de fallas, los modelos se deben definir de una forma diferente, de acuerdo con la topología del sistema. Las zonas se definen en relación a los nodos que poseen más de una fase en el sistema. Se debe descartar los sectores con configuraciones monofásicas. Esto significa que no todos los nodos estarán definidos dentro de las zonas de las propuestas de acuerdo a las figuras 7.16, 7.17 y 7.18. Incluso en algunos modelos se descartará el uso de algunas zonas dentro del análisis de falla por ser zonas que contienen nodos con conexiones monofásicas. La tabla 7.23 presenta los resultados de las clasificaciones durante la aplicación de los modelos con las 3 propuestas de distribución de zonas dentro del sistema.

Tabla 7.23: Porcentaje de acierto de modelos propuestos para fallas bifásicas línea-línea nivel III

Propuesta	**No.**	**Porcentaje por grupo** [%]				
	Zonas	**Zona1**	**Zona2**	**Zona3**	**Zona4**	**Zona5**
1	3	92,22 %	93,30 %	92,44 %	-	-
2	4	92,33 %	93,50 %	-	-	-
3	5	92,77 %	93,10 %	92,44 %	-	-

Los modelos de las propuestas 1 y 3 obtienen mejores estimaciones en comparación con los modelos de la propuesta 2. La diferencia fundamental entre las propuestas radica en una mejor clasificación de las observaciones para fallas

ubicadas en las zonas ramales del sistema para la propuesta 1, a diferencia de la propuesta 3 que realiza mejores estimaciones en la parte principal del sistema. Por otro lado si analizamos la distribución de nodos dentro de las zonas de acuerdo a las propuestas y el tipo de falla; en la propuesta 3 se distribuyen equilibradamente el número de nodos disponibles para análisis de falla. En la propuesta 1, la primera zona posee un número mayor de nodos en contraste con las otras zonas (la segunda zona sólo posee un nodo disponible). Lo anterior permite identificar a la propuesta 3 como un clasificador más informativo que el presentado por la propuesta 1.

Fallas bifásicas doble línea a tierra

En este tipo de falla se utiliza la misma definición de nodos y zonas utilizadas para las fallas bifásicas línea-línea. En este tipo de fallas los porcentajes de estimación obtenidos están por debajo de los porcentajes de otros tipos de falla según se ilustra en la tabla 7.24

Tabla 7.24: Porcentaje de acierto de modelos propuestos para fallas bifásicas doble línea a tierra nivel III

Propuesta	**No.**	**Porcentaje por grupo** [%]				
	Zonas	**Zona1**	**Zona2**	**Zona3**	**Zona4**	**Zona4**
1	3	88,22 %	75,33 %	82,22 %	-	-
2	4	88,88 %	77,77 %	-	-	-
3	5	89,20 %	87,77 %	83,33 %	-	-

La mayoría de las deficiencias en las estimaciones del *nivel III* de clasificación, están condicionadas en falsas estimaciones del *nivel II* de clasificación. La propuesta 3 presenta los mejores resultados en la estimación correcta de los datos de validación. Los modelos de la propuesta 2 presentaron buenos resultados individuales. Sin embargo su poca descriptibilidad y su bajo porcentaje de estimación lo hacen incompetente para ser utilizado como clasificador de observaciones de falla para observaciones bifásicas doble línea a tierra.

Fallas trifásicas

Al igual que las fallas bifásicas, los modelos generados para la clasificación de las fallas trifásicas tienen entre dos y tres áreas de influencia debido a la configuración monofásica en dos ramales del sistema, en los cuales no aplica el análisis de fallas trifásicas. La tabla 7.25 expresa los porcentajes de acierto alcanzados por cada propuesta para este tipo de falla.

Tabla 7.25: Porcentaje de acierto de modelos propuestos para fallas trifásicas nivel III

Propuesta	**No.**	**Porcentaje por grupo** [%]				
	Zonas	**Zona1**	**Zona2**	**Zona3**	**Zona4**	**Zona4**
1	3	97,33 %	98,02 %	99,30 %	-	-
2	4	92,30 %	97,50 %	-	-	-
3	5	99,20 %	98,11 %	98,11 %	-	-

En la zona 1 se presentaron porcentajes menores de aciertos en su mayoría, durante la clasificación de las observaciones de la barra 12. Sin embargo este fenómeno se presentó durante la clasificación de observaciones con valores menores de resistencia de falla (10Ω y 20Ω). La mayor parte de los errores en las estimaciones se atribuyen en problemas de estimación en el *nivel II* de clasificación. En este sector del sistema, los modelos de la propuesta 2 presentaron más inconvenientes para clasificar correctamente las observaciones. Probablemente la zona de influencia de las observaciones del nodo 12 sea similar a las observaciones del nodo número 7. Es posible que se presente una falta de sensibilidad del modelo ya que ambos nodos se encuentran en los límites de congruencia de los *clusters* de las dos zonas, produciendo un conflicto a la hora de clasificar adecuadamente estas observaciones. En los reportes generados se presenta entre un 32.1 % y 49.3 % de probabilidad de asignación en la zona 1 de las observaciones del nodo 12, contra un 50,8 % y 67.8 % de probabilidad de asignación de estas observaciones en la zona 2 del modelo. Lo que explica los resultados obtenidos.

Aunque los modelos de las propuestas 1 y 3 presentan resultados similares de eficiencia en las estimaciones, la propuesta 3 presenta una descripción más detallada de las zonas debido a sus distribuciones más pequeñas que encie-

rra pocos nodos. La propuesta 1 por otro lado define una cantidad mayor de nodos para la zona 1, por lo que las interpretaciones de las estimaciones son más generalizadas.

7.5.2.3. Clasificación conjunta según el *nivel II* (rangos de resistencias de falla) y el *nivel III* (zonas de nodos)

En las secciones anteriores se presentaron por separado los resultados obtenidos para las propuestas definidas para el *nivel II* y el *nivel III* de clasificación. En esta sección se presenta la comparación de los 12 modelos generados a partir de la combinación de las propuestas de los niveles *II* y *III* para construir clasificadores de localización de fallas. Con cada tipo de falla se generaron los parámetros para 12 modelos diferentes durante la etapa de clasificación.

Cada modelo se etiquetó de acuerdo al tipo de propuesta utilizada en los dos últimos niveles para identificarlos entre sí. La forma de identificación corresponde a la asignación de una letra **R** seguida de un número, para identificar el tipo de propuesta correspondiente al *nivel II* de clasificación (rangos de resistencia de falla). Seguido se identifica la propuesta del *nivel III* (zonas de nodos) con la letra **Z**, acompañada de su número correspondiente. Por ejemplo, el modelo construido a partir de la propuesta 1 del *nivel II* y la propuesta 3 del nivel *III*, se designa por su etiqueta $\mathbf{R}_1\mathbf{Z}_3$.

Fallas monofásicas

Los resultados de las clasificaciones correctas realizadas a las 180 observaciones de fallas monofásicas disponibles para validación, presentaron mejores estimativos para los modelos $\mathbf{R}_1\mathbf{Z}_2$,$\mathbf{R}_1\mathbf{Z}_3$ y $\mathbf{R}_2\mathbf{Z}_3$, presentados en la tabla 7.26.

Los resultados muestran mejores estimaciones con modelos que utilizan la propuesta 1 para el *nivel II* de clasificación, representados por los modelos $\mathbf{R}_1\mathbf{Z}_2$ y $\mathbf{R}_1\mathbf{Z}_3$. Estos modelos poseen altos porcentajes de estimación para observaciones con valores de resistencia altos y una efectividad menor con

Tabla 7.26: Porcentaje de acierto de los modelos generados para fallas monofásicas

R	**Z**	**Probabilidad por grupo** [%]			
		10Ω	**20Ω**	**30Ω**	**40Ω**
1	1	73,33	88,89	95,56	91,11
1	2	75,50	86,67	93,33	95,56
1	3	84,44	82,22	93,33	100
2	1	33,33	80,00	93,33	77,78
2	2	83,33	77,78	91,11	84,44
2	3	93,33	82,22	91,11	93,33
3	1	80,00	80,00	93,33	80,00
3	2	73,33	82,22	91,11	88,89
3	3	82,22	82,22	91,11	93,33
4	1	77,78	77,78	86,67	86,67
4	2	73,33	80,00	84,44	84,44
4	3	80,00	86,67	84,44	91,11

las observaciones de falla que tienen valores menores de resistencia de falla. A diferencia de estos modelos, el modelo $\mathbf{R}_2\mathbf{Z}_3$ es un poco más uniforme en los porcentajes de acierto para las estimaciones a lo largo de los rangos evaluados. El error en las estimaciones se presentó principalmente en la zona media del sistema. La mayor parte del error durante la estimación fue debido a problemas con la clasificación adecuada de las observaciones dentro de los rangos correctos de resistencia de falla para los nodos 14, 15, 18, 19 y 20. Si se comparan los modelos correspondientes a la misma propuesta de rangos de resistencia, se observará que los modelos definidos bajo la propuesta de 5 zonas para el *nivel III* de clasificación, registraron una cantidad mayor de aciertos en las estimaciones realizadas. Las estimaciones realizadas por los criterios *BIC*, *ICL* y *AIC* concuerdan parcialmente con la evaluación de los modelos como clasificadores. Según los criterios de selección, los modelos con cuatro zonas deben ser los más acertados en las estimaciones seguido de los modelos con cinco zonas definidas. La afinidad de los dos resultados se nota en los porcentajes obtenidos por los modelos $\mathbf{R}_1\mathbf{Z}_2$,$\mathbf{R}_1\mathbf{Z}_3$ y $\mathbf{R}_2\mathbf{Z}_3$.

Fallas bifásicas Línea - Línea

Los resultados obtenido de las estimaciones de 144 observaciones de validación para fallas bifásicas a tierra presentados en la tabla 7.27, muestra un comportamiento bastante homogéneo de cada uno de los 12 modelos generados.

Tabla 7.27: Porcentaje de acierto de los modelos generados para fallas bifásicas línea-línea

R	**Z**	**Probabilidad por grupo** [%]			
		10Ω	**20Ω**	**30Ω**	**40Ω**
1	1	99,99	99,90	99,99	99,99
1	2	94,44	94,44	99,90	99,90
1	3	99,90	99,99	99,99	99,90
2	1	99,90	99,99	99,99	99,99
2	2	83,33	83,33	94,44	95,57
2	3	99,99	99,99	99,99	99,99
3	1	94,44	99,99	99,99	99,99
3	2	94,44	94,44	94,44	94,44
3	3	99,70	99,95	99,90	99,95
4	1	99,98	99,98	99,99	99,99
4	2	94,44	94,44	94,44	95,57
4	3	99,90	99,99	99,90	99,99

Aunque las estimaciones realizadas presentan altos porcentajes de acierto en la clasificación de los datos; los modelos con cuatro zonas definidas para el *nivel III* de clasificación presentaron bajos porcentajes de acierto en relación con las estimaciones obtenidas con otros modelos. Ante la similaridad de los resultados, la selección de modelos se establece con base en la capacidad de obtener una información más detallada de las clasificaciones. En este caso el grado de detalle se puede encontrar en los modelos que definan un número mayor de zonas, con lo cual el número de zonas por grupo disminuye. Este aspecto favorece una localización más puntual de la falla durante la inspección en sitio del sistema. Los modelos $\mathbf{R}_3\mathbf{Z}_3$ y $\mathbf{R}_4\mathbf{Z}_3$ pueden seleccionarse como clasificadores de este tipo de fallas para el sistema estudiado.

Fallas bifásicas doble línea a tierra

Los porcentajes de eficacia de las estimaciones realizadas a 144 observaciones de fallas bifásicas doble línea a tierra, fueron menores que los alcanzados con los modelos de falla bifásica línea-línea, según se aprecia en los resultados de la tabla 7.28. En su mayoría los errores en las estimaciones fueron detectados en la clasificación de los nodos 7, 8 y 12 del sistema.

Tabla 7.28: Porcentaje de acierto de los modelos generados para fallas bifásicas doble línea a tierra

R	**Z**	**Probabilidad por grupo** [%]			
		10Ω	**20Ω**	**30Ω**	**40Ω**
1	1	66,66	86,11	86,11	86,90
1	2	80,55	91,66	83,33	63,88
1	3	86,11	86,11	100	94,44
2	1	63,88	82,22	86,11	99,90
2	2	83,33	83,33	94,44	94,44
2	3	83,33	91,66	97,22	94,44
3	1	77,77	83,33	86,11	86,11
3	2	99,99	91,66	91,66	91,66
3	3	99,88	91,66	77,77	94,44
4	1	75,00	80,00	83,33	72,22
4	2	91,66	91,66	91,66	91,66
4	3	72,22	83,33	97,22	94,44

Para este tipo de fallas los porcentajes de acierto fueron muy variados entre un modelo y otro. En general se observa un mejor comportamiento de los modelos con números grandes de grupos para los niveles *II* y *III* de clasificación. Destacándose los modelos $\mathbf{R}_3\mathbf{Z}_2$,$\mathbf{R}_3\mathbf{Z}_3$ y $\mathbf{R}_4\mathbf{Z}_2$. Aunque el modelo $\mathbf{R}_1\mathbf{Z}_3$ ofrece muy buenos resultados, presenta deficiencias en la clasificación de observaciones con valores menores de resistencia de falla. Según los criterios *BIC*, *ICL* y *AIC*, los modelos más representativos de la distribución de datos evaluada durante el entrenamiento, corresponde a aquellos que definen dos zonas dentro del sistema estudiado de acuerdo a la propuesta 2 en el *nivel III* de clasificación (ver figura 7.17). A diferencia de la propuesta 1 que también

define tres grupos dentro del sistema para este tipo de fallas, la propuesta 2 separa en grupo diferentes los nodos 7 y 12 del sistema. Con lo anterior se consigue diferenciar a estos nodos durante el cálculo de parámetros y se consiguen estimaciones correctas de sus posiciones durante la clasificación. La propuesta 3 del *nivel III*, toma esta idea y aumenta en tres zonas la distribución de nodos. El resultado final un mejor porcentaje de estimaciones en para los modelos que utilizan la propuesta como se puede observar en los modelos $\mathbf{R}_1\mathbf{Z}_3$,$\mathbf{R}_2\mathbf{Z}_3$, $\mathbf{R}_3\mathbf{Z}_3$ y $\mathbf{R}_4\mathbf{Z}_3$. Sin embargo estos modelos no son lo suficientemente eficientes tal y como ocurre con los modelos $\mathbf{R}_3\mathbf{Z}_2$ y $\mathbf{R}_4\mathbf{Z}_2$, con resultados muy superiores. En los modelos $\mathbf{R}_3\mathbf{Z}_2$ y $\mathbf{R}_4\mathbf{Z}_2$ se sacrifica detalle para una mejor eficiencia en las estimaciones.

Fallas trifásicas

En la tabla 7.29, se describen los estimados de clasificaciones correctas de los modelos generados, sobre las 96 observaciones de fallas trifásicas disponibles. Los resultados fueron en su totalidad próximos al ciento por ciento esperado.

Tabla 7.29: Porcentaje de acierto de los modelos generados para fallas trifásicas

R	**Z**	**Probabilidad por grupo** [%]			
		10Ω	**20Ω**	**30Ω**	**40Ω**
1	1	99,90	100	99,50	99,90
1	2	99,99	99,88	99,77	100
1	3	99,88	99,90	100	99,80
2	1	99,80	99,97	100	99,89
2	2	92,22	92,40	93,50	100
2	3	99,88	99,90	99,70	99,80
3	1	99,90	99,99	99,88	99,50
3	2	92,30	92,31	92,77	99,99
3	3	99,80	99,95	100	99,80
4	1	99,80	99,99	99,89	99,80
4	2	91,99	93,88	92,90	99,99
4	3	99,88	99,90	100	99,80

Los modelos que incluyen la propuesta 2 para el *nivel III* de clasificación presentan menores porcentajes de acierto en la clasificación de los datos de validación. Mediante el análisis dato por dato de las estimaciones realizadas, la mayoría de los errores de clasificación ocurren en las observaciones de falla de los nodos 7 y 12 del sistema. Estos nodos se encuentran en la zona límite de las dos zonas definidas por la propuesta 2 para fallas trifásicas. El resto de los modelos propuestos mostraron resultados de estimación muy altos, por lo que cualquiera podría ser igual de válido para ser utilizado como clasificador del sistema estudiado. Sin embargo debido a una distribución más detallada de los rangos y las zonas dentro del sistema, sería muy útil usar el modelo $\mathbf{R}_4\mathbf{Z}_3$, que presentaría información con más detalle de la localización de las fallas registradas. En estas condiciones, sólo el analista decide el grado de detalle requerido para sus estimaciones.

Capítulo 8

Conclusiones y trabajos futuros

8.1. Conclusiones

A través del procedimiento desarrollado, se destaca la capacidad de utilizar una pequeña cantidad de descriptores para generar modelos de gran precisión en sus estimaciones.

El uso de los componentes principales como herramienta estadística, permite utilizar los métodos de inferencia con mayor rapidez y procesos de cálculo más simplificados. Al reducir las dimensiones de los datos utilizados, el número de parámetros calculados se reduce, traducido en menores espacios de memoria ocupados dentro del ordenador.

La propuesta desarrollada, se convierte en una herramienta alternativa de fácil acceso y operación para localización de fallas en sistemas de distribución. El manejo del paquete no requiere conocimientos avanzados en sistemas de potencia, a diferencia de otros programas basados en métodos clásicos de análisis de falla.

El desarrollo de este tipo de alternativas favorece a las empresas distribuidoras y operadoras de red, que necesitan mantener y mejorar los índices de calidad del servicio a través de la disminución de los tiempos de atención de fallas.

El uso de herramientas basada sobre análisis estadístico, requiere de sistemas confiables para adquisición de datos. El éxito de los modelos durante las estimaciones, depende de la calidad en la obtención y procesamiento de la información recopilada. Lo anterior compromete a las empresas en la adquisición de sistemas de información confiables, como complemento para estimaciones eficientes.

Las representaciones DF y los métodos gráficos descritos son alternativas que facilitan el análisis de los datos durante la etapa de reconocimiento de patrones. A través del despliegue de elementos gráficos, el analista obtiene rápidamente información relevante durante el análisis de conglomerados.

El empleo de los criterios *BIC*, *ICL* y *AIC* establecen un marco de referencia sobre la cantidad óptima de grupos para construir los modelos clasificadores. Su aplicación se enfoca en la generación de suficientes componentes de distribución. Al final, el analista es quien tiene en sus manos, la decisión sobre la cantidad de grupos que desea generar.

La definición de modelos adecuados requiere tiempo para comparar todas las posibles alternativas. El analista tiene la responsabilidad de verificar las relaciones de las observaciones disponibles, así como las condiciones iniciales de operación de los algoritmos. Lo anterior permite modelos capaces de realizar estimaciones con un alto nivel de eficiencia.

La metodología propuesta se basa en la implementación de procesos de rápida respuesta, aplicables a procedimientos de atención de fallas y recuperación del sistema, que minimicen los tiempos de interrupción del servicio.

El autor expresa su total satisfacción durante el desarrollo de este trabajo, como integrante del grupo de investigación *GISEL*. El conocimiento adquirido durante este período, permite destacar el gran potencial de desarrollo generado por parte de sus investigadores. La capacidad de liderazgo y el nivel de

conocimiento hace posible la creación de alternativas prácticas y novedosas en las líneas de investigación en calidad de suministro de energía eléctrica, monitoreo de la continuidad del suministro e índices de calidad del servicio.

8.2. Recomendaciones

Antes de utilizar los modelos generados como clasificadores oficiales, se debe realizar múltiples pruebas entre modelos que presentan mejor respuesta.

Para el correcto análisis de sistemas de distribución, se necesita una base de datos con gran cantidad de observaciones de falla en todos los puntos del sistema, con el fin de representar el comportamiento del sistema bajo condiciones de falla.

Las zonas del *nivel III* de clasificación deben contener como mínimo tres nodos para evitar la no convergencia de los métodos durante la etapa de entrenamiento.

Para caracterizar un sistema de distribución muy grande, es útil dividirlo por sectores para estimaciones más claras.

La etapa de entrenamiento es diferente en cada sistema analizado. El uso de modelos calculados para otros sistemas puede inducir a errores de estimación.

8.3. Trabajos futuros

La siguiente etapa supone la evaluación del desempeño de modelos generados con datos provenientes de sistemas reales de distribución, bajo condiciones de falla.

Se deja abierta la posibilidad de incorporar nuevos descriptores que permitan definir alternativas en la metodología, para incrementar el nivel de precisión de los modelos.

Se propone la implementación de etapas intermedias que incorporen técnicas como las redes neuronales, que mejore la construcción de modelos durante la etapa de entrenamiento.

Se puede mejorar los algoritmos desarrollados hacia la construcción de modelos con mezclas adaptativas. A través de las mezclas adaptativas, los parámetros de los modelos son actualizados ante la presencia de nuevos estadísticos en la base de datos. Esto ocurre sin necesidad de correr nuevamente los algoritmos de generación de mezclas. Lo anterior supone generar modelos más dinámicos, acordes al crecimiento constante de los sistemas de distribución.

Las técnicas utilizadas pueden enfocarse hacia la prevención de eventos, mediante la estimación de zonas con mayor probabilidad de riesgo de falla dentro de los sistemas. Lo anterior puede enfocarse hacia la disminución de la duración y frecuencia de las interrupciones.

Apéndice A

Manual de operación del paquete estadístico

A.1. Introducción

Las funciones que contienen los algoritmos de estimación de fallas para sistemas de distribución mediante las técnicas estadísticas estudiadas se encuentran implementadas dentro del paquete llamado *MF_prog*. La carpeta llamada *MF_prog* debe cargarse al path de directorios de MATLAB para su ejecución. Adicionalmente se debe añadir al path, la carpeta compstats, que contiene una serie de comandos de aplicaciones estadísticas creados por MathWorks [Martínez *et al*, 2002]. Para desplegar la interfaz del paquete, se debe ejecutar el comando *MFa* en el *prompt* de la pantalla principal. (figura A.1)

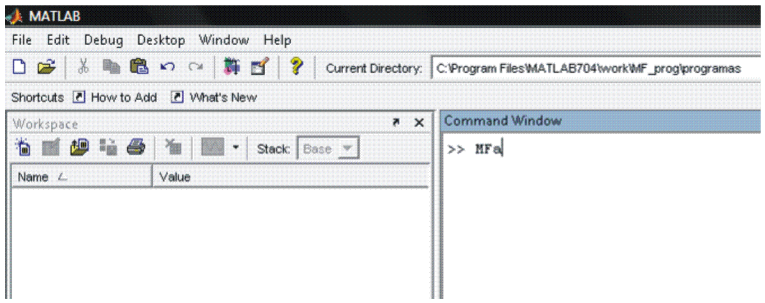

Figura A.1: Acceso al paquete a través del *prompt* en MATLAB

En la ventana de la interfaz (figura A.2), se encuentran distribuidos una serie de botones y campos, útiles para desarrollar cada una de las etapas de entre-

namiento, análisis y clasificación de los datos disponibles para el analista.

1. Botón de apertura de banco de observaciones de falla.
2. Botón de selección de modelos clasificadores.
3. Botón para ejecutar procesos de entrenamiento o de clasificación.
4. Campos de selección tipos de falla en estudio.
5. Campos de selección etapa de entrenamiento y etapa de clasificación.
6. Botón para selección de centros a través de algoritmo k-means.
7. Botón representación DF de clusters por zonas (nivel III).
8. Botón representación DF de clusters por rangos de resistencia de falla (nivel II).
9. Botón despliegue de ventana para revisión de criterios BIC, ICL y AIC.
10. Botón despliegue de reporte de clasificación de observaciones de falla.
11. Cuadro de observaciones de ayuda.

La ventana contiene un cuadro de ayuda adicional que describe brevemente las acciones ejecutadas al seleccionar o pulsar cada uno de los campos y botones existentes. De igual forma, cada botón posee mensajes de ayuda tipo *tip-string*[1] que permiten conocer las funciones ejecutadas por de la interfaz al pulsarlos (figura A.3).

A.2. Etapa de entrenamiento

Los datos destinados al entrenamiento del paquete informático, deben organizarse de forma particular, para que éstos sean leídos adecuadamente. El paquete utilizará los datos de las señales de tensión y corriente para procesarlos y generar los parámetros de los modelos necesarios para la evaluación

[1]Estos mensajes aparecen cuando se ubica el cursor por encima de los botones o campos de la interfaz.

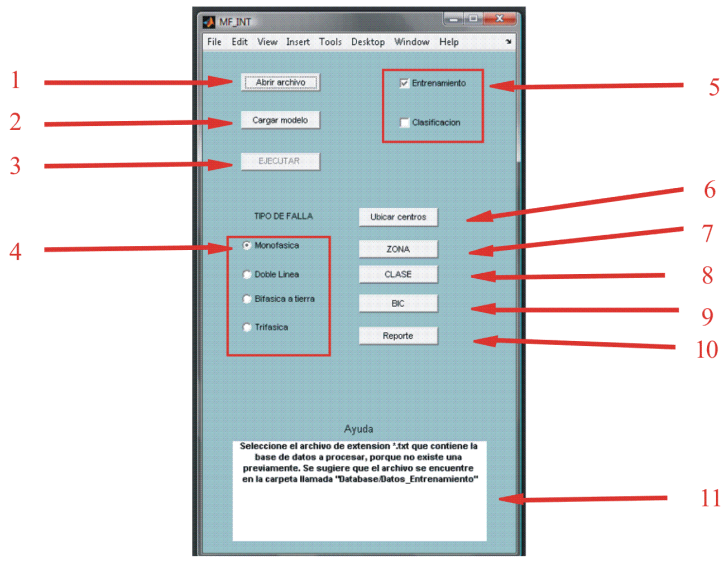

Figura A.2: Distribución de botones y campos de la interfaz

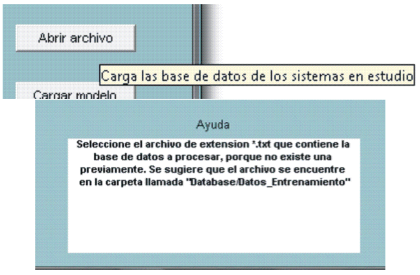

Figura A.3: Cuadro de Ayuda con información de cada función de la interfaz y mensajes tipo *tip-string* que posee cada botón

y clasificación de nuevos datos de falla provenientes del sistema estudiado.

Los datos deben organizarse de acuerdo al tipo de falla dentro del sistema (falla monofásica, bifásica línea-línea, bifásica doble línea a tierra, trifásica) en archivos independientes. En cada archivo los datos deben organizarse de la siguiente manera.

Enumeración de los nodos

Cada nodo de los sistemas estudiados debe enumerarse previamente para establecer una serie de etiquetas que permita la identificación de la posición de

las perturbaciones debidas a las fallas dentro de cada sistema.

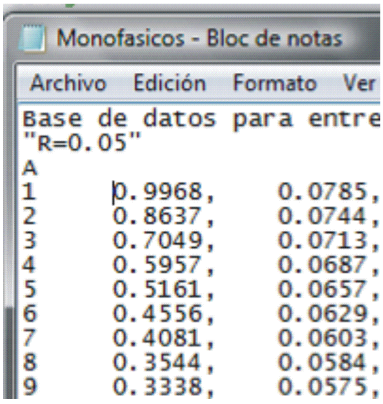

Figura A.4: Enumeración de los nodos dentro del archivo con los datos de entrenamiento

Organización de los datos

Los datos destinados a la etapa de entrenamiento se organizan dentro de una matriz donde cada fila de la matriz representa una observación de una falla registrada dentro del sistema, almacenados en archivos de texto plano (*.txt*). Cada fila contiene 6 columnas que representarán los descriptores utilizados para calcular los parámetros de cada modelo generado y una columna adicional de enumeración de nodos (figura A.5). El orden de los descriptores consignado en cada columna es el siguiente:

1. Número del nodo fallado.
2. Magnitud del hueco de tensión fase A (h_{Va})
3. Magnitud del hueco de tensión fase B (h_{Vb})
4. Magnitud del hueco de tensión fase C (h_{Vc})
5. Magnitud del pico de corriente fase A (h_{Ia})
6. Magnitud del pico de corriente fase B (h_{Ib})
7. Magnitud del pico de corriente fase C (h_{Ic})

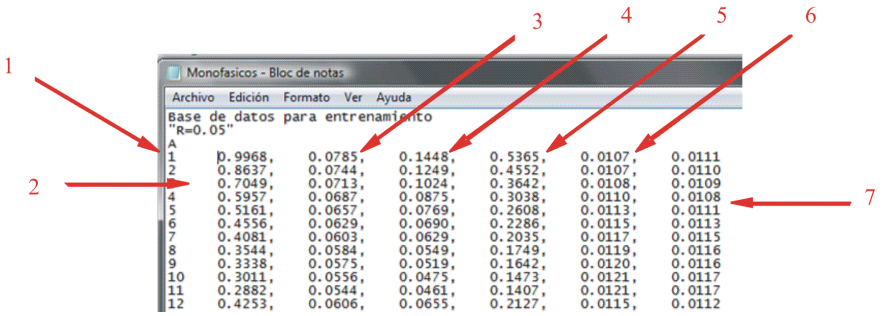

Figura A.5: Distribución de la información de los datos de entrenamiento dentro del archivo de lectura del sistema

Entre más datos existan para la etapa de entrenamiento, mejor serán los resultados obtenidos durante el proceso de cálculo de los parámetros que describen las distribuciones de los *clusters* o conglomerados dentro de la distribución muestral presentada.

El archivo de datos para entrenamiento debe poseer el título "Base de datos para entrenamiento". Los datos deben organizarse en paquetes que representan las observaciones de acuerdo al valor asociado de resistencia de falla registrado (ver figura A.5). Para diferenciar entre cada paquete se debe incluir la etiqueta, que represente el valor de resistencia de falla asociado encerrando el valor (por ejemplo: "$R = 26$"). El orden de ingreso dentro de los paquetes no se rige por la numeración de los nodos dentro del sistema. Pueden ingresarse múltiples observaciones de falla del mismo nodo con igual valor de resistencia de falla asociado sin importar el orden dentro del paquete. Los archivos con las observaciones de entrenamiento deben almenarse dentro de la carpeta con ruta:

C:\... \ work \ MF_prog \ Database \ Datos_Entrenamiento

Como se menciona en la sección 6, se utilizará un entrenamiento semi-supervisado para generar los modelos de mezclas. Lo anterior significa que tanto el número, como la forma de las distribuciones de los clusters, serán generados de acuerdo a las condiciones iniciales acotadas por el analista. La forma de definir los modelos está basada en el número de grupos presente en

la mezcla, para las zonas de ubicación y los rangos de resistencia de falla. Para definir la manera de generar los modelos se emplea un archivo de texto plano (.txt) adicional (figura A.6), en el cual se consigna la forma de organizar los valores de resistencia de falla de acuerdo a los rangos sugeridos, y las barras del sistema dentro del tipo y número de zonas especificadas.

Los archivos que definen la forma de los modelos, se pueden utilizar para generar modelos a diferentes sistemas que poseen información similar. La información de los archivos de construcción de modelos debe ser compatible con los datos de entrenamiento del sistema sometido a estudio. Lo anterior hace referencia a que los valores de resistencia de falla y la cantidad de nodos en ambos archivos deben coincidir, de lo contrario el procesamiento de la información será erróneo y los modelos no serán generados apropiadamente. Las implicaciones anteriores, permiten que no sea necesario generar un nuevo archivo asociado para construcción de modelos cada vez que ingresamos información de nuevos sistemas, siempre y cuando se cumpla las compatibilidades descritas.

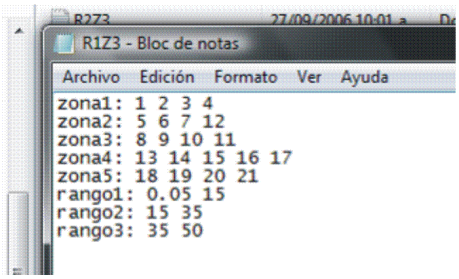

Figura A.6: Esquema en archivo .txt para definición de rangos y zonas

Los archivos de construcción de modelos deben archivarse en la carpeta denominada Desc_modelos ubicada en la carpeta work de Matlab, en la ruta:

C:\... \ work \ MF_prog \ Database \ Datos_modelos

Desde esta carpeta el sistema se encargará de cargar el archivo seleccionado para ser utilizado en la etapa de entrenamiento para efectos de generación de

los modelos estadísticos.

Procedimiento de operación de la interfaz bajo la etapa de entrenamiento

En la fase de entrenamiento, se toma el archivo con los datos de falla del sistema y mediante la selección del archivo de construcción de modelos, se genera el modelo estadístico acorde a las características del sistema estudiado. El procedimiento para realizar este proceso es el siguiente:

- Una vez ubicados en la ventana de la interfaz *MF_INT*, se selecciona la opción *entrenamiento* y el tipo de falla en estudio (figura A.7).

figura A.7

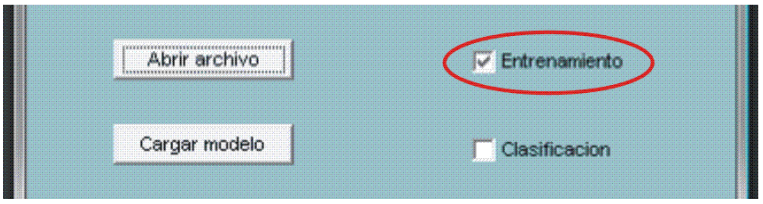

Figura A.7: Campo para selección de etapa de entrenamiento de la interfaz

- Cargar el archivo con los datos de falla, mediante el botón *abrir archivo* (figura A.8). Automáticamente aparecerá el cuadro en el cual se encuentran las carpetas con los archivos de falla. Seleccionamos el archivo correspondiente al sistema en estudio y ejecutamos la opción *abrir*.

- Una vez seleccionado el archivo, se pulsa la opción *cargar modelo* y se selecciona el tipo de modelo que se desea generar, de acuerdo a las características del sistema y la lista disponible de archivos para construcción de modelos (figura A.9). Hay que considerar la compatibilidad de los archivos para seguir con el proceso.

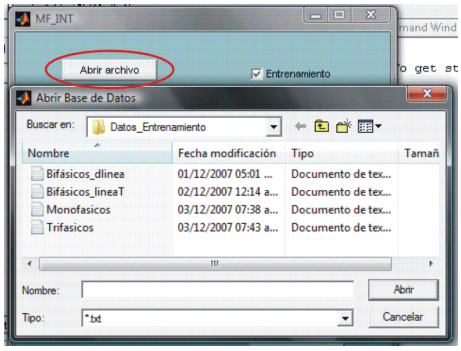

Figura A.8: Botón para apertura de archivo bases de datos entrenamiento

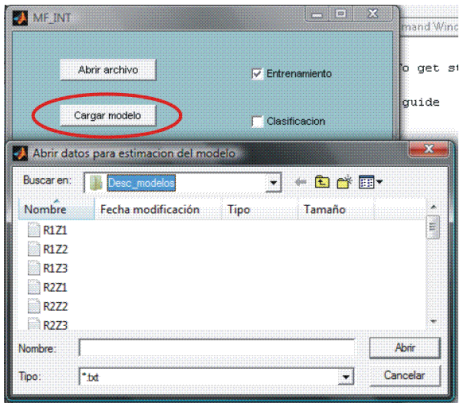

Figura A.9: Botón para cargar modelo de entrenamiento

- Mediante el botón *ubicar centros* se inicia la búsqueda de centros de los grupos de la primera etapa del modelo por medio del algoritmo *EM*. Al pulsar el botón, se despliega una pantalla con la distribución de los datos y los centros estimados aparecen como puntos de color rojo (figura A.10). Es posible cambiar el punto de visión de la distribución pulsando el botón izquierdo del mouse en la ventana, y rotando la imagen de la distribución. Si los centros estimados no son del agrado del analista, es posible estimar nuevos centros pulsando nuevamente el botón *ubicar centros*.

- Cuando se obtienen los centros deseados se pulsa el botón *EJECUTAR*. Esta acción toma toda la información que se ha ingresado, se procesa y el modelo estadístico es generado (figura A.11).

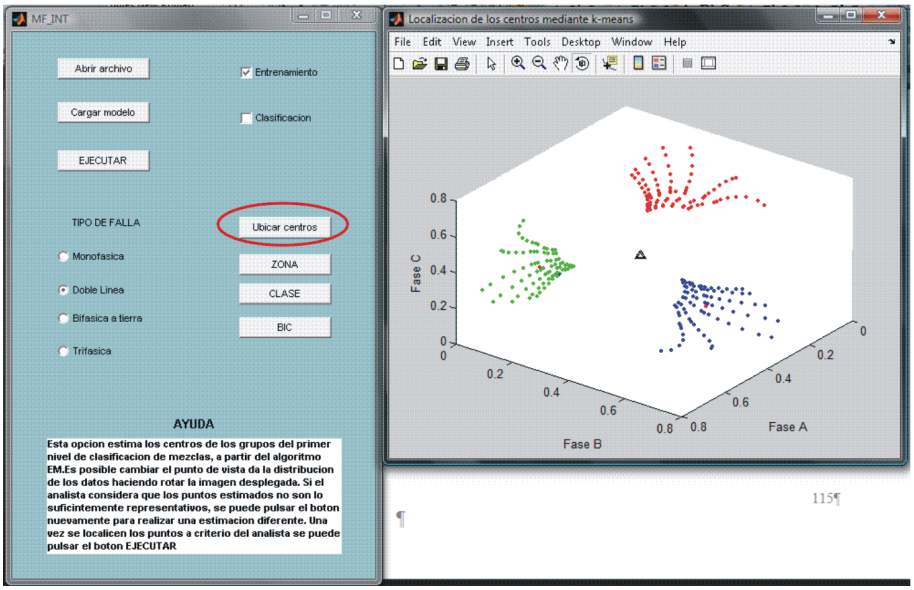

Figura A.10: Selección de centros iniciales por medio del algoritmo k-means

- Una vez generado el modelo, éste debe ser salvado bajo un nombre. Una ventana se despliega automáticamente para guardar el modelo generado según el tipo de falla seleccionado (figura A.12).

A.3. Etapa de clasificación

Datos de clasificación

Los datos destinados para clasificación, consiste en información preparada para determinar la ubicación de fallas de acuerdo a parámetros almacenados en modelos de caracterización generados previamente. Los datos de clasificación deben estar organizados en archivos de texto plano (*.txt*) para poder procesarse dentro del sistema de la interfaz.

La forma de ingresar los datos de validación dentro de los archivos de texto plano, obedece un orden similar con el cual se organiza la información en la fase de entrenamiento. La matriz de datos contiene filas con los valores de tensión y corriente de las tres fases bajo condiciones de falla. El archivo de

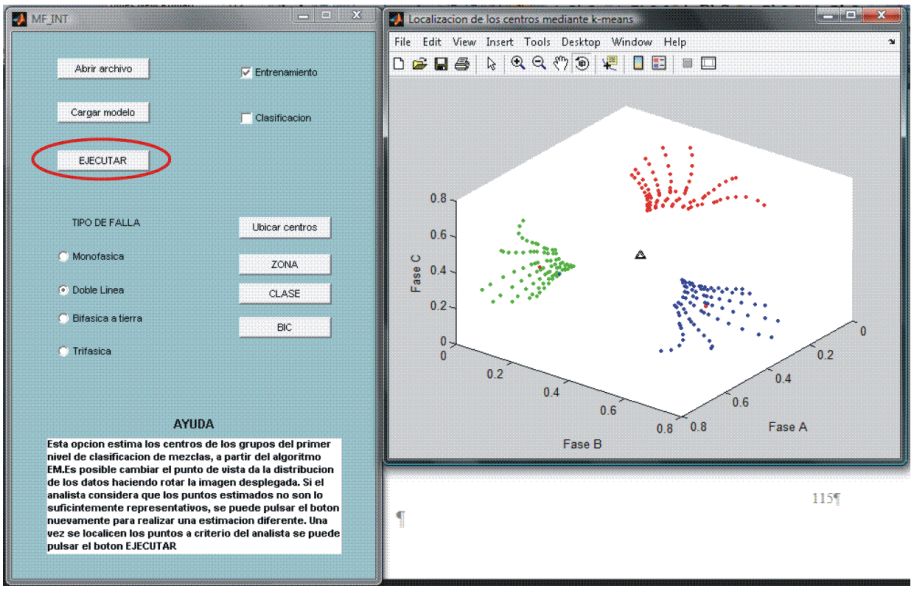

Figura A.11: Ejecución de etapa de entrenamiento para generar parámetros de los modelos

datos debe iniciar con el título "Datos para clasificación". En este tipo de archivos las observaciones de falla no se rotulan con el número del nodo fallado, y no existe un orden de ingreso de cada observación dentro de la matriz de datos, debido a la naturaleza desconocida de la ubicación de estas observaciones dentro del sistema analizado (figura A.13).

Dentro del paquete de la interfaz existen habilitadas unas carpetas destinadas a almacenar los datos de clasificación de acuerdo al tipo de falla registrado. Los archivos con la información para efectos de clasificación debe archivarse en la carpeta denominada Datos_evaluacion ubicado en la ruta:
C:\... \ work\ MF_prog \ Database

Procedimiento de operación de la interfaz bajo la etapa de clasificación

En la etapa de clasificación, se utilizan los modelos generados para estimar la posible pertenencia de los datos disponibles en los grupos definidos dentro de las mezclas realizadas. El procedimiento a seguir es el siguiente:

- Se cargan los datos disponibles para efectos de clasificación. Se selecciona

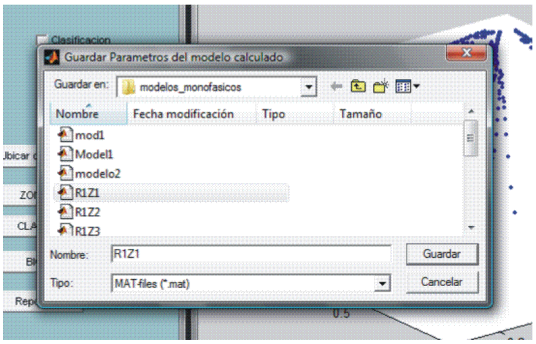

Figura A.12: Ventana para salvar los parámetros calculados durante etapa de entrenamiento

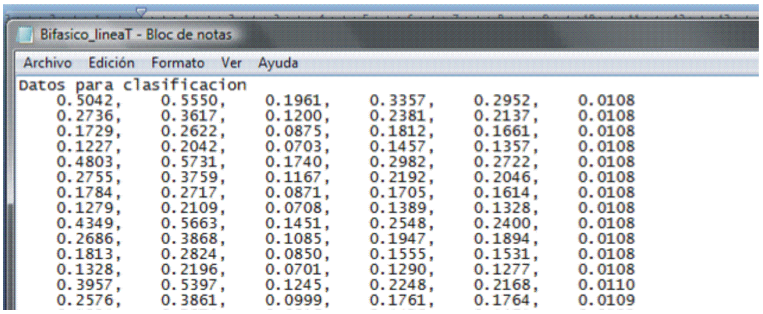

Figura A.13: Archivo de almacenamiento de observaciones para localización

la opción *clasificación*, se especifica el tipo de falla en estudio, y luego se pulsa el botón *abrir archivo*. El programa automáticamente abrirá la carpeta donde se encuentran los archivos destinados para ser clasificados. (Figura A.14)

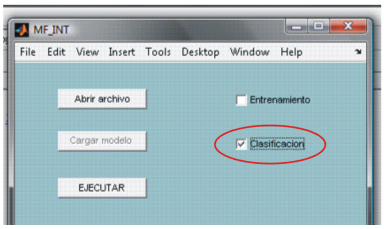

Figura A.14: Campo de selección de etapa de clasificación

- Se debe seccionar el tipo de modelo a ser utilizado en la clasificación. Esto se realiza con la opción cargar modelo (figura A.15).

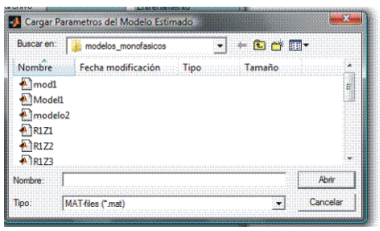

Figura A.15: Cuadro para selección de modelo clasificador

- Al oprimir *EJECUTAR*, el programa realizará el proceso de clasificación de acuerdo al modelo escogido y la información producida se almacenará en un archivo, el cual debe ser salvado bajo un nombre. Para conocer el resultado de la clasificación se debe pulsar el botón *reporte* y automáticamente se abrirá una ventana con el contenido del resultado de la clasificación (figura A.16).

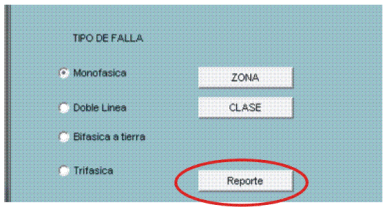

Figura A.16: Botón para generar el reporte de clasificación de datos

En el reporte generado aparece una serie de campos con la siguiente información:

1. Los valores de los rangos definidos para el modelo.
2. El número de zonas definidas y los nodos presentes en cada zona.
3. La clasificación de cada observación según el *nivel I*.
4. La clasificación de cada observación según el *nive II*.
5. La clasificación de cada observación según el *nive III*.

En cada nivel de clasificación aparece información relacionada de las observaciones junto con el grupo en el cual la observación fue clasificada y la

probabilidad de pertenencia de las observaciones dentro de cada grupo. En la figura A.17, se puede apreciar la distribución de los campos con la información relacionada con el reporte de clasificación de datos.

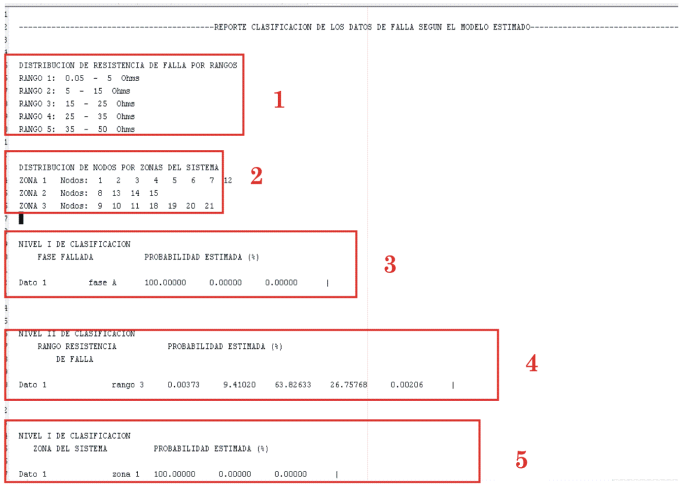

Figura A.17: Cuadro de reporte de clasificación de observaciones

A.4. Comparación de los modelos generados

Los modelos generados en la etapa de entrenamiento pueden ser comparados con otros modelos para la misma distribución de datos. Esto permite establecer los modelos que mejor describan el sistema estudiado, aplicando el método de máxima verosimilitud.

Mediante el empleo de los criterios BIC, ICL y AIC (ver capítulo 5), es posible obtener una apreciación inicial de lo cercano que está cada modelo de describir los datos utilizados en el entrenamiento.

Una vez realizada la etapa de entrenamiento y generados diversos modelos para un sistema en particular, se puede utilizar la función BIC. Esta función despliega una ventana adicional (figura A.18), la cual permite seleccionar hasta seis modelos diferentes, todos generados a partir de la misma base de datos. El procedimiento para utilizar esta opción es el siguiente:

Figura A.18: Ventana de aplicación de criterios BIC, ICL y AIC

- Utilizando la pestaña desplegable ubicada en la parte superior de la ventana (figura A.19), se puede seleccionar el número de modelos a comparar. Automáticamente aparecerá una serie de campos, determinados según el número de modelos seleccionados en la pestaña superior de la ventana (figura A.20). Los campos que aparecen en la ventana son habilitados para escoger uno a uno los modelos que van a ser analizados por los criterios *BIC*, *ICL* y *AIC*.

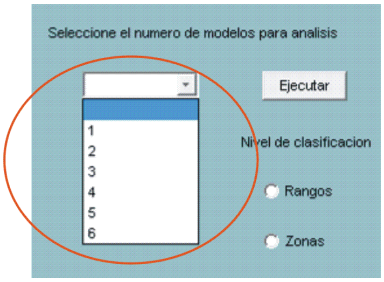

Figura A.19: Pestaña de selección de números de modelos a comparar por los criterios BIC, ICL y AIC

En esta ventana existe la opción de comparar los modelos según los rangos de resistencias de falla, o por el número de zonas que se hayan estipulado en cada uno de los modelos. Este tipo de selecciones están ubicados en la parte derecha de la ventana y puede utilizarse una a la vez (figura A.20).

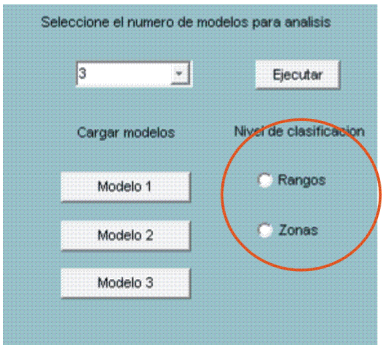

Figura A.20: Botones para selección de modelos y campos de selección de tipo de nivel de clasificación

- Seleccionados las condiciones de número de modelos y la manera de comparar los modelos, se procede a pulsar el botón *EJECUTAR*. Automáticamente aparece una ventana adicional con la representación grafica de los valores calculados por los criterios para cada modelo, según el número de grupos que cada uno de ellos genera. Dentro del grafico, aparecerán tres curvas diferentes que representan a cada criterio respectivamente (figura A.21).

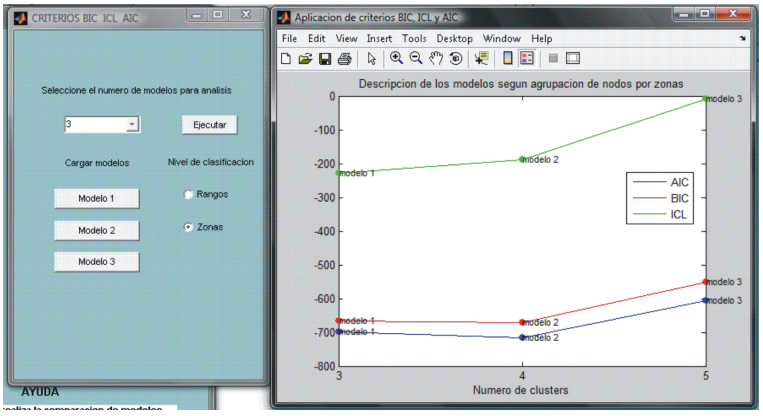

Figura A.21: Despliegue de representación gráfica de índices calculados por los criterios BIC, ICL y AIC

Para comprender la representación gráfica desplegada, basta con observar dentro de cada curva, los índices derivados del cálculo de cada criterio que posean el menor valor registrado. En la figura A.21, el menor valor registrado

para la curva que representa el criterio *BIC* (curva roja), corresponde al modelo compuesto por 4 grupos cuyo índice calculado corresponde a −1100. Según el criterio *BIC*, este modelo se postula como el más indicado para ser utilizado como clasificador de los datos de falla suministrados. El uso de este gráfico permite analizar los modelos entre sí y realizar un contraste entre los resultados de cada criterio.

A.5. Distribución de los datos dentro de los clusters generados mediante representaciones gráficas DF

El empleo de representaciones gráficas DF (*Distribution Functions*) permite desplegar de forma conjunta, una imagen de la distribución de los datos disponibles y las representaciones de cada uno de los grupos generados mediante los parámetros de las mezclas de distribuciones de probabilidad. Mediante las representaciones DF, es posible apreciar la posición de cada dato disponible en relación con los dominios de cada función de distribución presente dentro de la mezcla generada. Entiéndase por dominio, la forma y tamaño de cada grupo dentro del espacio muestral, de acuerdo a los parámetros de la función de probabilidad asignada.

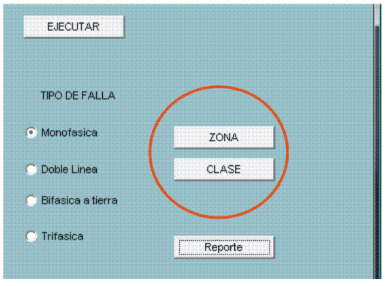

Figura A.22: Botones para generar representaciones DF por clases (*nivel II*) y por zonas (*nivel III*)

A través de las opciones *zonas* y *clase* es posible generar este tipo de gráficos sobre una serie de ventanas adicionales que se despliegan automáticamente (figura A.22). Esta opción puede aplicarse en las dos etapas de la interfaz.

Durante la etapa de entrenamiento, permite percibir el tamaño y forma de cada grupo generado por los parámetros de los modelos utilizados. En la etapa de clasificación, sirve de ayuda para observar la posición de los datos clasificados respecto a cada uno de los grupos disponibles por el modelo utilizado.

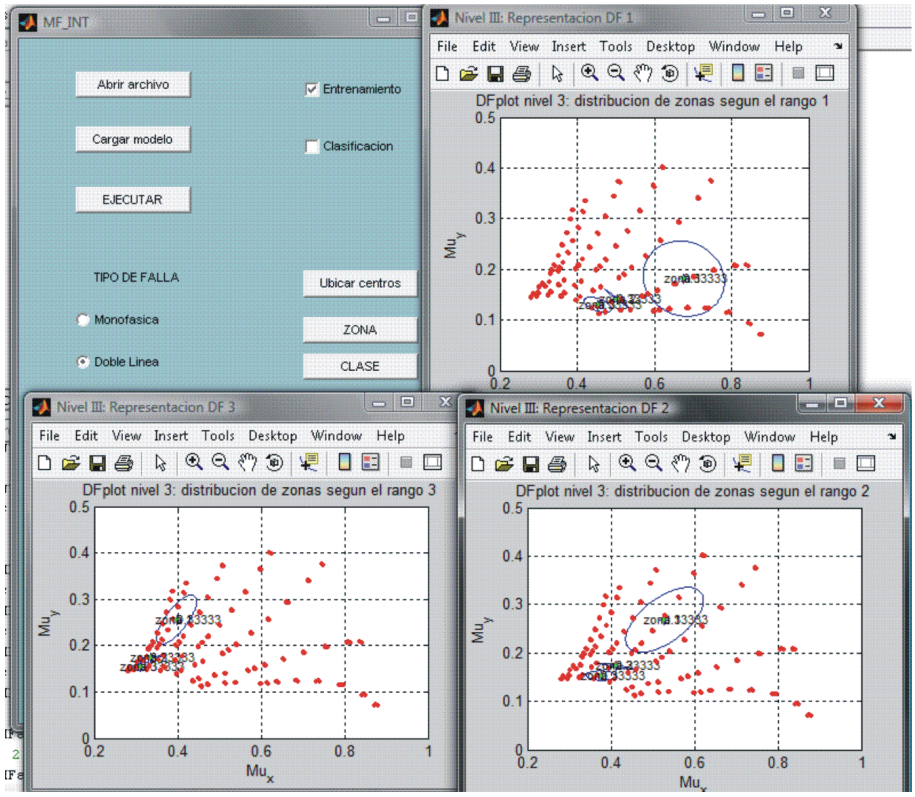

Figura A.23: Ventanas de representaciones gráficas DF de localización de clusters en la distribución de datos seleccionados (*nivel III* de clasificación)

Al ejecutar la función *zonas*, se presentarán una serie de ventanas. Cada ventana presenta un gráfico DF referente a cada rango definido de resistencia de falla en el cual aparece la distribución de las zonas definidas (figura A.23). Esto significa que por cada rango definido de resistencia de falla, existe una distribución diferente de las zonas de agrupación de las barras el cual es presentado. Estas representaciones DF corresponde con la distribución de las zonas y las observaciones utilizadas en el espacio muestral, de acuerdo a cada rango de resistencia de falla generados en el *nivel II* de clasificación del

sistema.

A través de la función clase, los grupos de clasificación por rangos de resistencia de falla se despliegan en una representación DF dentro de una ventana independiente (figura A.24). Lo anterior significa que existe una ventana con la representación DF de las observaciones y la distribución de los grupos representativos de los rangos de resistencia de falla determinados.

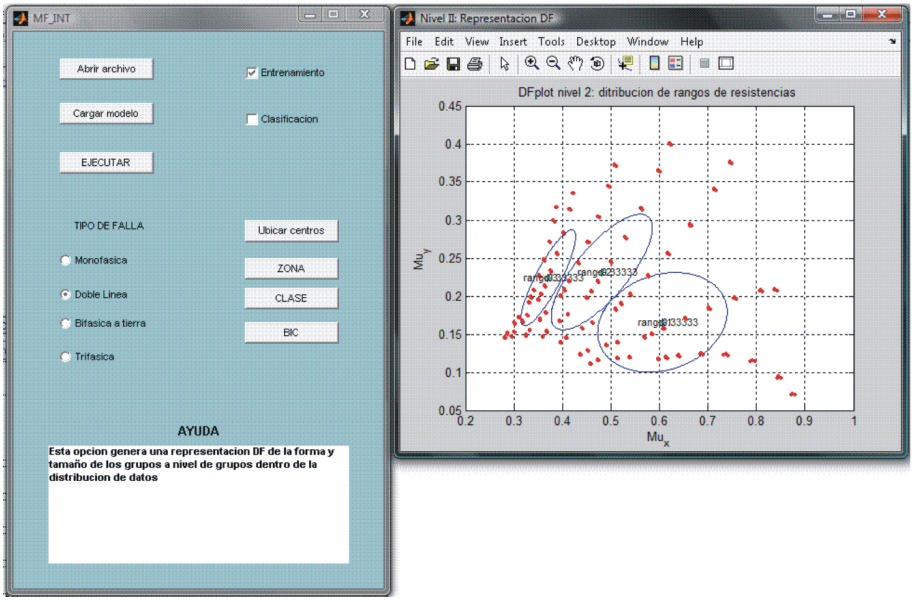

Figura A.24: Ventana de representación gráfica DF de localización de clusters en la distribución de datos seleccionados (*nivel II* de clasificación)

En las representaciones DF aparecerá información adicional sobre valores de las probabilidades *a priori* de cada distribución y la identificación del número de zona o rango de correspondiente.

Apéndice B

Otros métodos estadísticos de clasificación

B.1. Introducción

En situaciones de discriminación o clasificación de datos, una vez se conoce los parámetros de las distribuciones, el problema admite una solución general. En la mayoría de los casos los parámetros son desconocidos y deben estimarse a través de los mismos datos. En muestras multivariantes de gran tamaño, el uso del cálculo de las distancia de Mahalanobis para distribuciones predominantemente normales es considerablemente aceptado [Peña, 2002]. Sin embargo es frecuente que los datos disponibles presenten distribuciones que no sean normales. Este es el caso de datos con problemas de clasificación con variables discretas.

Existen métodos alternativos, unos basados en la posibilidad de construir modelos que expliquen los valores de cada variable, utilizando modelos de respuesta cualitativa. Por otro lado existen métodos que requieren el uso intensivo del computador y se fundamentan en algoritmos y funciones convenientemente construidas.

B.2. Árboles de clasificación

Los árboles de clasificación (*Classification And Regression Trees*, CART) no utilizan un modelo estadístico formal y es más bien un algoritmo para clasificar, utilizando particiones binarias sucesivas de una variable cada vez [Martínez et al, 2002]. Si se supone que existe una muestra de entrenamiento con información de los grupos de pertenencia de los datos, que permita construir el modelo de clasificación. Es posible aplicar este criterio para clasificar nuevos datos. El proceso se inicia creando un nodo inicial basado en una pregunta capaz de dividir el conjunto disponible de datos en dos grupos homogéneos utilizando una de las variables. El algoritmo selecciona una de las variables x_i, y se obtiene un punto de corte c, de manera que se puede separar los datos tales que $x_i < c$, de aquellos condicionados por $x_i > c$. A partir de este nodo inicial se generan dos nodos más, a donde llegarán los datos discriminados en el primer nodo. En cada nodo se repetirá el proceso de selección de una variable y un punto de corte para dividir los datos en dos partes homogéneas. El proceso termina cuando se clasifican todas o el mayor número de observaciones correctamente en su grupo. La construcción del árbol requiere de las siguientes decisiones:

- La selección de las variables y del punto de corte.
- Cuándo un nodo se considera terminal y cuándo no.
- La asignación de clases a los nodos terminales.

Para decidir la variable a utilizar para realizar la partición en el nodo, primero se calcula la proporción de observaciones que pasan por el nodo. Para cada uno de los grupos, se utiliza una medida llamada entropía (B.1), la cual mide la impureza del nodo t según la probabilidad de que las observaciones que pasan por este nodo pertenezcan a cada una de las clases en las cuales el punto de corte separa los datos $p(g|t)$. La entropía se define como:

$$I(t) = -\sum_{g=1}^{G} p(g|t) \log p(g|t) \tag{B.1}$$

La variable utilizada para realizar la división de los datos en un nodo, se selecciona minimizando la heterogeneidad o impureza resultante de la división. El proceso de separación de los datos puede representarse gráficamente de acuerdo a la figura B.1.

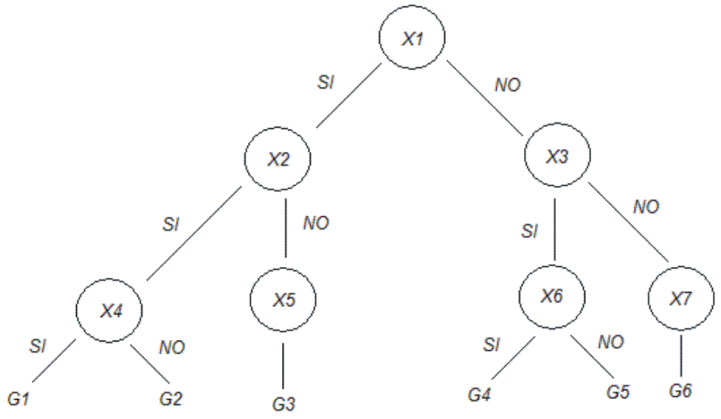

Figura B.1: Representación gráfica de un árbol de clasificación

La clasificación de los nodos terminales se hace asignando todas las observaciones del nodo, al grupo más probable en ese nodo. Si la impureza del nodo es cero, todas las observaciones pertenecen al mismo grupo y su clasificación puede hacerse sin error. En el caso contrario existe un cierto error en la clasificación si la impureza del nodo no es cero. El proceso de construcción del árbol puede generar muchos nodos si el número de variables es grande, lo que hace necesario realizar procesos para simplificar o "podar el árbol" para hacerlo más manejable sin pérdida importante de información.

B.3. Redes neuronales

Las redes neuronales son algoritmos de análisis de datos, basado en el uso intensivo del computador. Su aplicación se apoya según el teorema de Kolmogorov (1957) [Peña, 2002], en que cualquier función continua de múltiples variables puede aproximarse como suma de funciones univariantes. La función multivariante continua f, con las variables x_i se puede aproximar a la función (B.2).

$$y = f(x_1, ..., x_p) = \sum_{i=1}^{N} g_i(z) \tag{B.2}$$

Donde las g_i son funciones continuas de una variable y N puede ser muy alto dependiendo del grado de precisión que se pretenda alcanzar. Las redes neuronales se construyen a partir de elementos llamados nodos, entradas o neuronas. Estas unidades reciben un conjunto de entradas representadas por la variable vectorial $\mathbf{x}$, y calculan una variable escalar de salida aplicando una ponderación a los componentes de entrada, añadiendo una constante de sesgo, y transformando el resultado de forma no lineal de acuerdo a la expresión (B.3).

$$\mathbf{z} = g(\mathbf{w}'\mathbf{x}) \tag{B.3}$$

Aunque existen muchas estructuras posibles de redes neurales, la más utilizada es el perceptrón, el cual consiste en un número de neuronas clasificadas en capas. En cada capa la variable de entrada en una de las neuronas, es la respuesta construida por combinación lineal de la variable de entrada $\mathbf{x}$ y la función de ponderación $\mathbf{w}$, proveniente de las neuronas de la capa inmediatamente anterior.

Para llevar a la práctica este método, es necesario estimar los parámetros que definen cada función g_i. Las redes neuronales necesitan muestras muy grandes para estimar eficazmente la gran cantidad de parámetros que requieren. La posibilidad de trabajar en problemas de altas dimensiones es una ventaja, ya que las variables $\mathbf{x}$ se sustituyen por las proyecciones $\mathbf{w}'\mathbf{x}$, para trabajar con funciones de una variable. Una red neuronal con una estructura bien diseñada puede dar resultados similares a los métodos clásicos de clasificación bajo condiciones estándar, y comportarse mejor en situaciones donde las relaciones entre las variables no sean lineales.

B.4. Máquinas de soporte vectorial

También llamadas *máquinas del vector soporte*, este método presenta un enfoque distinto del habitual. En lugar de buscar una reducción del espacio de los datos y resolver el problema en una menor dimensión, se busca un espacio con mayor dimensión donde los datos puedan separarse de forma lineal. Si se dispone de una muestra con observaciones multivariantes tal que $x_i \in \mathbf{R}^P$, este conjunto de datos es linealmente separable, si es posible encontrar un vector $\mathbf{w} \in \mathbf{R}^P$, que defina un plano que separe correctamente las observaciones. Es decir, todas las observaciones que pertenezcan a un grupo se encontrarán de un lado del plano de separación, mientras las demás observaciones se encontrarán del lado opuesto del plano [Peña, 2002].

Si se tiene el valor del hiperplano de separación óptima entre dos conjuntos, dado por:

$$f(x_i) = \mathbf{w}'x_i + b \tag{B.4}$$

La distancia entre un punto cualquiera x_i, y el hiperplano dado por (B.4) será la proyección del punto x_i en la dirección $\mathbf{w}$ que es el vector ortogonal al plano. La proyección de un vector sobre otro se expresa como:

$$\frac{\mathbf{w}'x_i}{|\mathbf{w}|} \tag{B.5}$$

Al evaluar las distancias de cada observación al plano mediante la proyección ortogonal (B.5), y se cumple que $y_i(\mathbf{w}'x_i + b) \geq c$, entonces estos datos serán clasificados como pertenecientes al grupo ubicado hacia el lado positivo del plano y el resto serán separados hacia el lado contrario del mismo, tal y como se representa en la figura B.2.
El enfoque del vector soporte es aplicar una transformación a los datos que lleve a un espacio de dimensión mucho mayor que p, y entonces aplicar una discriminación lineal.

Para trasladar los puntos de un espacio de dimensión mayor, se introducen

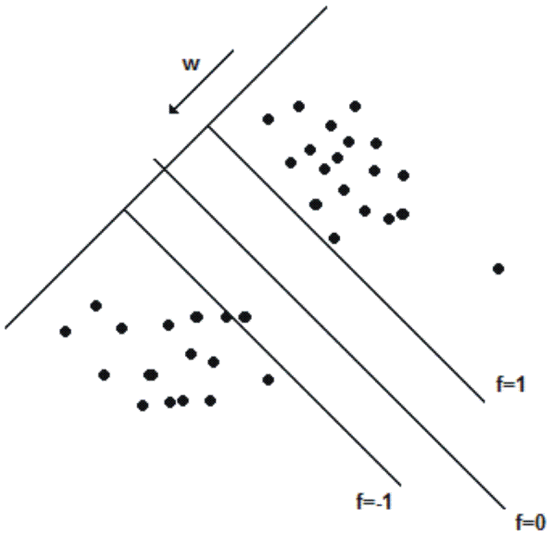

Figura B.2: Representación gráfica de dos clases separables linealmente, el plano separador **f** y el vector **w** ortogonal al plano separador

nuevas variables que sean potencias de las variables actuales, o en su lugar, productos de potencias de las variables iniciales. La clave consiste en conocer los productos escalares entre las observaciones dentro del espacio ampliado para resolver el problema.

Apéndice C

Análisis multidimensional

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad. Por tanto, si fuera posible describir con precisión los valores de p variables por un subconjunto formado por r variables, tal que $r < p$, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.

El análisis de componentes principales tiene como objeto, analizar a través de n observaciones con p variables, si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. La técnica de componentes principales es establecida por Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por Karl Pearson (1901).

La técnica parte de encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él las observaciones, éstas conserven su estructura con la menor dispersión posible. Si se supone que se dispone de los valores de p-variables en n elementos de una población dispuestos en una matriz $\mathbf{X}$ de dimensión n x p. Si a cada variable de la matriz se le ha restado su media, de manera que cada elemento de $\mathbf{X}$ tenga media cero y su matriz de covarianzas esté dada por $1/n$ $\mathbf{X}'\mathbf{X}$. El primer componente principal se definirá como combinación lineal de las variables originales que tienen varianza máxima

[Peña, 2002].

Los valores en este primer componente de n individuos se representará por un nuevo vector $\mathbf{z}_1$, dado por la expresión (C.1).

$$\mathbf{z}_1 = \mathbf{X}\, a_1 \tag{C.1}$$

Como las variables originales tienen media cero, también $\mathbf{z}_1$ tendrá media nula. Por lo tanto su varianza será:

$$\frac{1}{n}\,\mathbf{z}_1'\mathbf{z}_1 = \frac{1}{n}\,a_1'\mathbf{X}'\mathbf{X}a_1 = a_1'\mathbf{S}a_1 \tag{C.2}$$

Donde $'\mathbf{S}$ es la matriz de varianzas y covarianzas de las observaciones. Para que la maximización de la ecuación (C.2) tenga solución, se debe imponer una restricción al módulo del vector a_1, tal que $a_1'a_1 = 1$. Esta restricción se impone mediante el multiplicador de Lagrange definido en (C.3).

$$M = a_1'\mathbf{S}a_1 - \lambda(a_1'a_1 - 1) \tag{C.3}$$

La expresión (C.3) se maximiza de la forma habitual derivando respecto a los componentes de a_1, e igualamos a cero para obtener la expresión (C.4).

$$\frac{\partial M}{\partial a_1} = 2\mathbf{S}a_1 - 2\lambda a_1 = 0 \tag{C.4}$$

Cuya solución es (C.5)

$$\mathbf{S}a_1 = \lambda a_1 \tag{C.5}$$

La expresión (C.5) implica que a1 es un vector propio de la matriz $\mathbf{S}$, y λ es su correspondiente valor propio. Para determinar qué valor propio de $\mathbf{S}$ es la solución de la ecuación (C.5) se multiplica por a_1' esta ecuación, para obtener la expresión (C.6).

$$a_1'\mathbf{S}a_1 = \lambda a_1'a_1 = \lambda \tag{C.6}$$

Por medio de (C.2) se concluye que λ es la varianza de $\mathbf{z}_1$. Como esta cantidad es la que se busca maximizar, λ será el mayor valor propio de la matriz $\mathbf{S}$.

Su vector asociado a_1, define los coeficientes de cada variable en el primer componente principal.

Análogamente el espacio de dimensión r que mejor representa a las observaciones viene definido por los vectores propios asociados a los r mayores valores propios de $\mathbf{S}$. Estas direcciones se denominan direcciones principales de los datos y las nuevas variables por ellas definidas son los componentes principales.

En general la matriz $\mathbf{X}$ y por tanto $\mathbf{S}$ tienen dimensión p, existiendo tantas componentes principales como variables que se obtendrán, calculando los valores propios o raíces características $(\lambda_1, ..., \lambda_p)$, de la matriz de varianza y covarianzas $\mathbf{S}$. Mediante la expresión (C.7).

$$|\mathbf{S} - \lambda\mathbf{I}| = 0 \tag{C.7}$$

Y sus vectores asociados por la expresión (C.8).

$$(\mathbf{S} - \lambda_i\mathbf{I})a_i = 0 \tag{C.8}$$

Los términos λ_i son reales, al ser $\mathbf{S}$ una matriz simétrica, y positiva, ya que $\mathbf{S}$ es definida positiva. Si λ_j y λ_h son dos raíces distintas, sus vectores asociados son ortogonales, por ser $\mathbf{S}$ simétrica. Si $\mathbf{S}$ fuera semidefinida positiva dentro de un rango $r < p$, lo que ocurriría si $p-r$ variables fuesen combinación lineal de las demás, entonces habría solamente r raíces características positivas y el resto serían ceros. Llamando $\mathbf{Z}$ la matriz cuyas columnas son los valores de p componentes en n individuos, estas nuevas variables estarán relacionadas con las originales a través de (C.9).

$$\mathbf{Z} = \mathbf{XA} \tag{C.9}$$

Donde $\mathbf{A}'\mathbf{A} = \mathbf{I}$

Calcular los componentes principales equivale a aplicar una transformación ortogonal $\mathbf{A}$, a las variables $\mathbf{X}$ (ejes originales) para obtener nuevas variables

Z, incorreladas entre sí.

Los componentes principales son nuevas variables con las siguientes propiedades:

- Conservan la variabilidad inicial: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.
- La proporción de la variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz **S**.
- Las covarianzas entre cada componente principal y las variables **X** vienen dadas por el producto de las coordenadas del vector propio que define el componente por su valor propio.
- Las correlaciones entre un componente principal y una variable **X** es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.
- Las r componentes principales ($r < p$) proporcionan la predicción lineal óptima con r variables del conjunto de variables **X**.
- Al estandarizar los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

El procedimiento presentado es un enfoque general del proceso real, derivado de aplicar este tipo de técnicas en el manejo de datos proporcionados a través del estudio de eventos aleatorios. Definir el número de componentes principales requiere de múltiples intentos, de modo que sea posible encontrar la proyección más conveniente, capaz de explicar la población con la mínima cantidad de componentes.

Apéndice D

Coordenadas paralelas: un método alternativo para exploración gráfica de datos multivariantes

El uso del sistema de coordenadas cartesianas, limita el uso de representaciones gráficas a tres dimensiones referidos a sus respectivos ejes ortogonales. Si en vez de esto, se dibujara los ejes de referencia en forma paralela uno respecto el otro, se podría observar una gran cantidad de ejes sobre la misma representación. Esta técnica fue desarrollada por Wegman (1985) como una manera de observar y analizar los datos multidimensionales y fue introducida posteriormente en el contexto de la geometría computacional por Inselberg (1986) [Martínez *et al*, 2002].

La técnica de coordenadas paralelas fue expandida y descrita sobre un arreglo estadístico por Wegman (1990). Wegman expuso una explicación rigurosa de las propiedades de las coordenadas paralelas como una transformación proyectiva, e ilustró sus bondades en relación con las representaciones de las coordenadas cartesianas ortogonales.

Un gráfico de coordenadas paralelas para una representación p-dimensional de datos se construye dibujando p líneas paralelas entre sí. Cada línea representa el eje de coordenadas para $x_1, x_2, ..., x_d$. Cada eje posee la misma orientación positiva que el eje x de coordenadas cartesianas. En algunas representaciones

los ejes son dibujados en dirección vertical paralelos al eje y de coordenadas cartesianas. Un punto cualquiera $C = (x_1, x_2, x_3, x_4)$ se representa tal como aparece en la figura D.1[1]

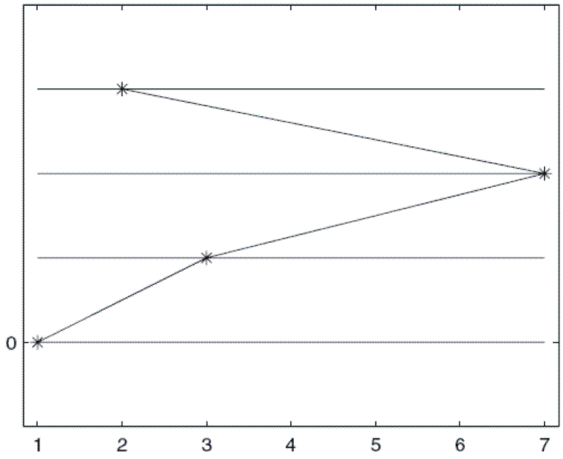

Figura D.1: Representación gráfica de un punto en el sistema de coordenadas paralelas

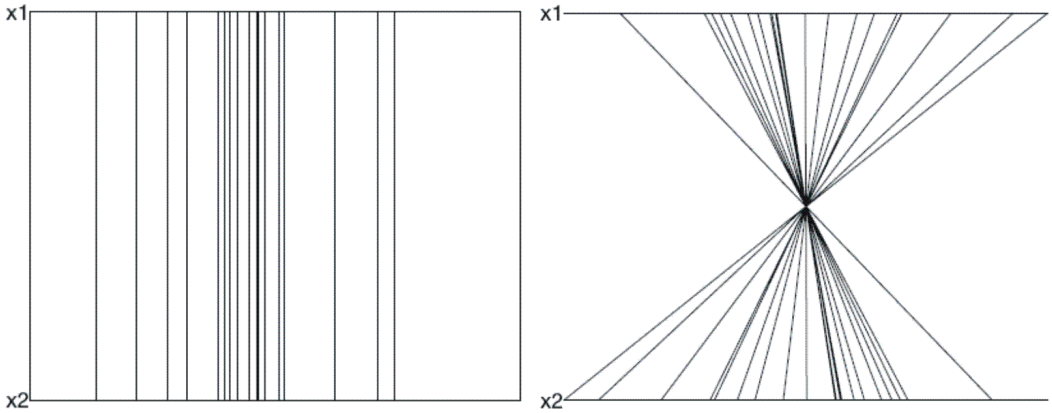

Figura D.2: Distribuciones de datos con índices de correlación 1 y -1 respectivamente, en el sistema de coordenadas paralelas

Las variables se ordenan hacia arriba y la representación de cada dato aparece como una línea poligonal con vértices localizados en cada eje de variables. Así cada punto representado en coordenadas cartesianas, aparece como una serie de segmentos de recta conectados en las representaciones de coordenadas

[1]Fuente: Computacional Statistics Handbook with MATLAB, Martínez Wendy, 2002

paralelas. Las representaciones en coordenadas paralelas permiten realizar una exploración de datos para indagar si las variables utilizadas son útiles para separar clases o no. Mediante el uso de coordenadas paralelas es posible medir el grado de correlación entre variables. En la figura D.2[2] aparece la representación de datos con índice de correlación de 1 y -1 respectivamente.

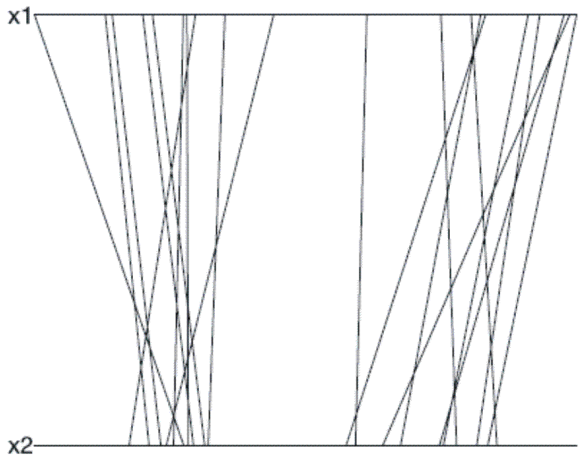

Figura D.3: Distribución de datos donde se evidencia la presencia de dos conglomerados

La existencia de conglomerados en una o más dimensiones puede ser identificada mediante la representación en coordenadas paralelas. La presencia de espacios o separaciones entre ejes y la convergencia de los segmentos hacia posiciones comunes son indicativos de la existencia de conglomerados, como aparece en las figuras D.3[3] y D.4[4].

La aplicación de las coordenadas paralelas es un intento de visualizar todas las observaciones disponibles y todas las dimensiones al mismo tiempo. La exploración gráfica de datos comprende generalmente una etapa en la cual se descubre qué pueden aportar los datos observados, sin utilizar estimaciones cuantitativas como errores de distribución, número de grupos y otros estadísticos numéricos. La exploración visual es una herramienta que ayuda a conformar un marco de referencia para entender mejor el modelo de los datos. Esta es una técnica complementaria que evita arrancar "a ciegas" en el

[2]Fuente: Computacional Statistics Handbook with MATLAB, Martínez Wendy, 2002
[3]Fuente: Computacional Statistics Handbook with MATLAB, Martínez Wendy, 2002
[4]Fuente: Computacional Statistics Handbook with MATLAB, Martínez Wendy, 2002

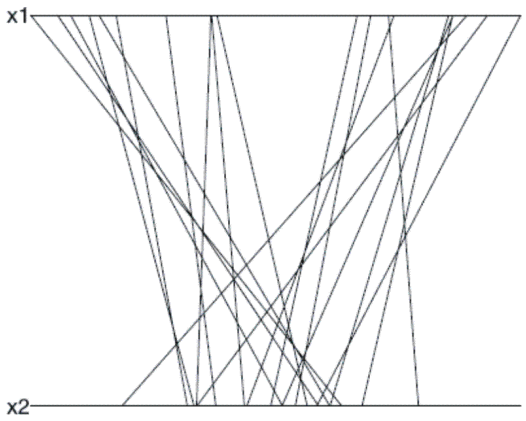

Figura D.4: Existencia de conglomerados en una sola dimensión

momento de aplicar otros métodos analíticos en las distribuciones estudiadas.

Bibliografía

Das R. *Determining the Locations of Faults in Distribution Systems*. PhD thesis,Saskatchewan University, Canada, 1998.

Dattatreya G. y Kanal L. "Estimation of Mixing Probabilities in Multiclass Finite Mixtures". IEEE transactions on systems, man, and cybernetics, Vol. 20, 1990.

Figueiredo M. "Unsupervised Learning of Finite Mixture Models". IEEE Transactions on pattern analysis and machine intelligence, Vol. 24, 2002.

IEEEstd1159-1995."IEEE Recommended practice for monitoring electrical power quality". 1995.

Martínez L. y Martínez A. *Computacional Statistics Handbook with MATLAB*. Chapman Hall, New York, 2002.

McLachlan G. y Peel D. *Finite Mixtures Models*. Willie, Canada, 2000.

Mora J. "Voltage Sag and Classification for Diagnosis in Electric Power Quality Domain". Technical Report, Department of Electronics, Computer Science and Autoatic Control-UDG, España, 2003.

Neimane V. On Development Planning of Electricity Distribution Networks. PhD Dissertation, Royal institute of Technology, Stockholm, 2001.

Paaβ G. y Kindermann J. "Bayesian Regresion Mixtures of Experts for Geo-References Data". Fraunhofer Institute for Autonomous Intelligent Systems-AIS, Germany, 2005.

Peña D. *Análisis de Datos Multivariantes*. McGraw-Hill, Madrid, 2002.

Priebe C., Rogers G., Marchette D. y Solka J. "Change Point Analysis with Adaptive Mixture Models". Center for computational statistics, George Mason University, Fairfax, 1993.

Rodríguez A. y Horst E. "Dynamic Density Estimation with Financial Applications". National Institute of Enviromental Health Science-NIH, Denmark, 2006.

Rubio J. "Experimentos Preliminares de Verificación de Locutores con una Base de Datos Realista". Escuela Técnica Superior de Ingenieros de Telecomunicación-ETSIT, 2000.

Sanjay S. Y Hebert T. "Bayesian Pixel Classification Using Spatially Variant Finites Mixtures and the Generalized EM algorithm". IEEE Transactions on image processing, Vol. 7, 1998.

Wang Y., Freedman M. y Kung S. "Probabilistic Principal Component Subspaces: A Hieralchical Finite Mixtures Models for Data Visualization". IEEE Transactions on neural networks, Vol. 11, 2000.

Yiming W., Xiangyu Y. y Kap Luk C. "Unsupervised Color Image Based on Gaussian Mixture Model". Nanyang Technological University, 2003.