



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA
ING. PETROLERA Y GAS NATURAL- EXPLORACIÓN

CLASIFICACIÓN AUTOMATIZADA DE FACIES A ESCALA DE REGISTROS DE
POZO USANDO REDES NEURONALES ARTIFICIALES EN YACIMIENTOS
ARENO – ARCILLOSOS TURBIDÍTICOS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN INGENIERÍA

PRESENTA:
CHRISTIAN AGNI RAMÍREZ HERRERA

TUTOR PRINCIPAL
MARTÍN A. DÍAZ VIERA
PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA

MÉXICO, D. F. ENERO 2014

JURADO ASIGNADO:

Presidente: Dr. Fernando Samaniego Verduzco
Secretario: Dr. Joaquín Eduardo Aguayo Camargo
Vocal: Dr. Martín Alberto Díaz Viera
1^{er.} Suplente: Dr. Guillermo Alejandro Pérez Cruz
2^{do.} Suplente: Dr. Ricardo Casar González

Lugar o lugares donde se realizó la tesis: MÉXICO D.F.

TUTOR DE TESIS:

Dr. Martín Alberto Díaz Viera



A handwritten signature in black ink, appearing to read 'Martín', is written over a horizontal dashed line. Below the line, the word 'FIRMA' is printed in a bold, sans-serif font.

FIRMA



Oficio

Remitente	Subdirección de Gestión de Recursos Técnicos Gerencia de Estrategias y Planes.	Fecha	30 de julio de 2013
Destinatario	Ing. Chistian Agni Ramírez Herrera Activo de Exploración Aguas Profundas Norte	Número	PEP-SGRT-GEP- 480 -2013
Asunto:	Autorización para utilización de información en tesis de maestría.	Antecedentes:	Número(s): Número único de expediente: Fecha(s):
		Anexo	<input type="checkbox"/>

En respuesta a la solicitud citada en antecedentes, sobre autorización para utilizar información técnica de pozos para la elaboración de la tesis de maestría, le informo que se considera procedente la petición para que sea incluida en dicha tesis, siempre y cuando se cumpla estrictamente con las políticas de confidencialidad sobre el uso de información vigentes en Pemex Exploración y Producción (no mencionar nombres específicos de yacimientos, instalaciones o pozos, así como información que se considere confidencial o estratégica reservas, localizaciones, coordenadas etc.), así como acatar el cumplimiento de acuerdos de confidencialidad sobre uso de información, que determine el Activo de Exploración Aguas Profundas Norte.

Aprovecho la ocasión para enviarle un cordial saludo.

Atentamente

Dr. Jorge A. Arévalo Villagrán
Gerente

C.c.p.- Dr. Pedro Silva López, Subdirector de Gestión de Recursos Técnicos.

Elaboró: mrhg



En memoria de David Sánchez López

El éxito es fácil de obtener. Lo difícil es merecerlo.

(Albert Camus)

Dedicatoria

A mis padres, Guadalupe Herrera Viveros y Juan Manuel Ramírez Jiménez quienes con su esfuerzo, dedicación y sacrificio han hecho posible que alcance triunfos y enfrente los fracasos. Gracias por todo el amor y el apoyo incondicional que me dieron a lo largo de mi vida.

A mis hermanos Nancy, Iván y Misael, que son mi fuente de inspiración y alegrías por ser el motor que me impulsa para seguir luchando por mis sueños

A Gaby, Lidia, Humberto, Jesús y Pamela, por todo su inmenso amor y todo el apoyo brindado a lo largo de todos estos años tanto en las buenas, como las malas. Gracias por ser el aliciente para seguir adelante.

A Olinca, Kenia, Iván, Leonel, Miguel, Mario y Eduardo V. por todo los buenos tiempos y su apoyo incondicional.

A Andrea y Valeria, por todo su amor y cariño, por llenarme de ánimo con su apoyo, en momentos difíciles, muchas gracias por llegar a mi vida.

A David Sánchez por todo el tiempo que mi familia y yo tuvimos la dicha de conocerte y compartir contigo algo más que amistad. Este logro es para ti, sé que algún día nos volveremos a encontrar.

Agradecimientos

La colaboración directa o indirecta de varias personas hizo posible la realización de este trabajo de Tesis, ya sea proporcionando información, guiando, leyendo, opinando, corrigiendo e impulsando a seguir adelante. Ofreciendo disculpas si algún nombre es involuntariamente omitido en estos agradecimientos.

Agradezco a mi asesor el Dr. Martín A. Díaz Viera por haber confiado en mi persona, por todo su apoyo, por sus consejos, por su paciencia, por transmisión de conocimientos y experiencia, pero sobre todo por su amistad.

Agradezco al Dr. Ricardo Casar González, por su amistad, por los datos, por todo el apoyo y las facilidades brindadas para que este trabajo fuera posible.

A la Ing. Paola Checa Rojas, a los C. Dr. Javier Méndez Venegas y Víctor Hernández Maldonado, por su apoyo, por el material y datos, por las asesorías y por su amistad.

Agradezco a los demás miembros del Jurado: al Dr. Fernando Samaniego Verduzco, al Dr. Joaquín Eduardo Aguayo Camargo y al Dr. Guillermo Alejandro Pérez Cruz, por todo su apoyo y sus valiosos comentarios

Un agradecimiento al Dr. Martín Cárdenas Soto y al M.C. David Escobedo Zenil, por todo el apoyo brindado en la gestión y trámites de posgrado.

Al Instituto Mexicano del Petróleo, por permitirme formar parte en el proyecto de investigación del cual salió este trabajo de Tesis, en especial al Grupo de Modelado Estocástico Espacial, que me cobijaron como integrante activo, el tiempo que tuve la dicha de participar en él. A todos los compañeros que conformamos este grupo gracias por transmitir sus conocimientos y por su apoyo.

A la Universidad Nacional Autónoma de México, por darme el honor de ser alumno de Posgrado, en su entonces programa de Maestría en Ingeniería Petrolera y Gas Natural. A todos mis profesores y a mis compañeros de maestría, Isabel, Alan y Javier, que junto a mi recorrieron este duro camino.

A Petróleos Mexicanos, por permitirme utilizar su información y darme las facilidades para poder concluir con este trabajo.

Agradecimientos Especiales

Al Dr. Jorge A. Arévalo Villagrán, Gerente de Estrategias y Planes, de la Subdirección de Recursos Técnicos, de PEMEX Exploración y Producción, por el apoyo y facilidades brindadas, para el uso y publicación de los datos en este trabajo de Tesis.

Al Ing. José Trinidad Martínez Vázquez, coordinador de Proyecto Área Perdido del Activo de Exploración Aguas Profundas Norte de la Subdirección de Exploración de PEP, por su apoyo, amistad y por las facilidades para concluir este trabajo.

A los Ingenieros Arturo Escamilla Herrera, Eduardo Macías Zamora, Carlos Alberto Caraveo Miranda, Carlos Rosas Lara, Rocío Negrete Cadena y Sergio Paz, por sus palabras de aliento y su apoyo.

A todos mis compañeros del proyecto Área Perdido, por su amistad y apoyo.

Abstract

The presence of a hydrocarbon reservoir in prospective areas depends on different geological factors; therefore the geologic framework plays an important role in different hydrocarbons exploration and production process. The geological conceptual model on an area allows the evaluation of the risk associated with petroleum system elements (hydrocarbon presence), reservoir size and profit. The depositional and facies models are very important to define a new play lead or reservoir; it allows inferences about the expected rock in a new well.

The well core data allow defining the depositional facies in a short interval of rock along the well; sometimes we don't have this information. For this reason we need to estimate depositional facies from indirect information of physical rock properties, like well logs or seismic. Hence we need to establish classification criteria for well logs data and associate the classes to different depositional system and/or kind of lithology. The statistical behavior of well logs data allows us to understand the nature of rock in the area. For this reason we use data exploratory analysis (DEA) and other statistical interpretation techniques and establish criteria to get classes called electrofacies.

The artificial neural networks (ANN) are computational systems to classify numeric information, frequently used to identify classes or electrofacies, in commercial software. The obtained results with neural networks depend on the nature of data, through the statistical knowledge, what it allows establishing training schemes to get optimum results. In this context, this thesis tries to establish a classification strategy by using well logs, with statistical analysis like main idea.

Resumen

La posible existencia de un yacimiento de hidrocarburos depende de los factores geológicos presentes en el área, por lo que el marco geológico de un área con atractivo económico petrolero es el elemento primordial, en muchas de las etapas de la exploración y explotación hidrocarburos. Los modelos geológicos conceptuales de un área, permiten evaluar los diferentes elementos de riesgo asociados a la presencia, el tamaño y la rentabilidad de un yacimiento posible. Es fundamental contar con un modelo adecuado de depósito y de facies; ya que permite definir un play, un prospecto o un posible yacimiento, para obtener inferencias sobre las características de los cuerpos rocosos que se esperan encontrar al perforar un pozo.

La información que mejor permite definir un modelo de facies depositacionales, son los núcleos de pozos, los cuales son solo un intervalo pequeño dentro del pozo y muchas veces no es posible contar con ellos, por lo que es necesario realizar inferencias a partir de información indirecta de las propiedades físicas, como lo es la información de registros de pozos y la sísmica. Es por lo anterior que existe la necesidad de establecer los criterios para clasificar esta información y así poder asociarla a diferentes sistemas de depósito y/o diferentes tipos de roca. Los registros geofísicos de pozo son datos cuantificables con un comportamiento estadístico que describe la naturaleza de los cuerpos rocosos presentes en un área. Es por ello que el análisis exploratorio de datos y otras herramientas estadísticas, permiten realizar interpretaciones y establecer los criterios para la clasificación de registros geofísicos de pozo en clases llamadas electrofacies.

Las redes neuronales artificiales (RNA) son sistemas computacionales que permiten la clasificación de cualquier tipo de información numérica; su aplicación en la clasificación de electrofacies comúnmente se implementa en softwares comerciales. Los resultados obtenidos mediante el entrenamiento de redes neuronales dependen de la naturaleza de los datos, por lo que el conocimiento de la información permite establecer esquemas de entrenamiento y con ello obtener resultados óptimos. En este contexto, este trabajo de tesis plantea una estrategia de clasificación de datos de registros de pozo mediante el uso de redes neuronales, fundamentado principalmente en el análisis estadístico.

ÍNDICE

INTRODUCCIÓN	1
I. LAS REDES NEURONALES ARTIFICIALES EN EL MARCO DE LA MODELACIÓN GEOLÓGICA PETROFÍSICA DE YACIMIENTOS (MGPY)	3
1.1 Modelación Geológica Petrofísica de Yacimientos (MGPY)	3
1.2 Componentes de la MGPY	3
1.3 Flujo de trabajo de la MGPY	4
1.4 Redes neuronales en la MGPY	7
II. REVISIÓN DEL ESTADO DEL ARTE DE LA APLICACIÓN DE LAS REDES NEURONALES ARTIFICIALES A LA CLASIFICACIÓN DE FACIES	9
2.1 Historia de las Redes Neuronales Artificiales.	9
2.1.1 Las RNA en la predicción de Litofacies	10
2.2 Las Redes Neuronales Artificiales (RNA)	12
2.2.1 Modelo Neuronal	13
2.2.2 Procesamiento y extracción de características.	16
2.3 Clasificación y tipos de redes Neuronales	16
2.3.1 De acuerdo con la naturaleza de los datos.	16
2.3.2 De Acuerdo con su topología.	17
2.3.2.1 Las redes Monocapa.	17
2.3.2.2 Las redes Multicapa.	18
2.3.3 Tipo de Aprendizaje.	19
2.3.3.1. Redes con aprendizaje supervisado	20
2.3.3.2. Redes con aprendizaje no supervisado	20
2.3.4 Tipo de asociación de las señales de entrada y salida	21
2.3.5 Por su aplicación	22
2.3.5.1 Reconocimiento de Patrones.	22
2.3.5.2 Aproximación de funciones	23
2.4 Arquitectura perceptrón.	23
2.5 Arquitectura ADALINE.	25

2.6 Red perceptrón multicapa.	26
2.6.1 Dilema de la Varianza y Bias	29
2.6.2 Ventajas y desventajas de una red MLP	31
2.7 Mapas auto-organizados	32
2.8 Sistema de múltiples redes	36
2.8.1 Combinación de conjunto	36
2.8.2 Red Modular (Modular Neural Network)	38
2.9 Regresión Lineal Multidimensional.	39
2.9.1 Analogía de la regresión lineal Multidimensional con las redes neuronales.	39
III. METODOLOGÍA	40
3.1 Censo y revisión de la información (control de calidad)	42
3.2 Análisis exploratorio de datos (AED).	42
3.2.1 Preparación de los datos	43
3.2.2 Análisis estadístico unidimensional y multidimensional	43
3.2.2.1 Variables categóricas.	44
3.2.2.2 Variables cuantitativas	44
3.2.2.3 Estadística univariada.	44
3.2.2.4 Estadística multivariada.	45
3.3 Análisis de Agrupamiento (AA) o Clústeres (AC).	46
3.3.1 Criterios de similitud.	47
3.3.1.1 Distancia euclidiana	47
3.3.1.2 Otros indicadores de distancia.	48
3.3.2 Criterios basados en medidas de correlación.	48
3.3.3 Similitud y distancia entre grupos	49
3.3.4 Algoritmos de análisis de grupos	49
3.3.4.1 Método de las k-medias	50

3.3.5 Análisis de agrupamiento en la clasificación de electrofacies.	51
3.4 Análisis de selección de variables.	52
3.4.1 Análisis de componentes principales (ACP).	52
3.4.1.1 Análisis de componentes principales en la estimación de electrofacies.	55
3.5 Diseño y entrenamiento de la red neuronal.	56
3.5.1. Topología.	56
3.5.2 Pesos y bias iniciales.	57
3.5.4 Funciones de activación	58
3.5.4 Datos de entrada	58
3.5.5 Algoritmos de entrenamiento	59
3.5.6 Entrenamiento	59
3.5.6.1 Entrenamiento de la red perceptron multicapas	59
3.5.6.2 Entrenamiento de los mapas auto-organizados.	60
3.5.6.3 Entrenamiento de las máquinas comité (Committee Machines)	60
3.6 Análisis del error y validación cruzada (durante la etapa de entrenamiento)	61
3.7 Clasificación multi-pozo de electrofacies mediante el modelo neuronal.	66
IV. ANÁLISIS DE LA APLICACIÓN DE REDES NEURONALES ARTIFICIALES A DATOS SINTÉTICOS	67
4.1 Registros geofísicos y criterio de clasificación	67
4.2 Registros sintéticos	69
4.3 Generación de las curvas sintéticas	69
4.3.1 Curva de rayos gama.	69
4.3.2 Curva de resistividad.	69
4.3.3 Curva Caliper	70
4.3.4 Curva de densidad	70
4.3.5 Curva de neutrón	70

4.3.6 Curva tiempo de tránsito (sónico)	71
4.3.7 Litotipos	71
4.3.8 Modelo uno (Bien clasificado)	72
4.3.9 Modelo dos (Medianamente clasificado)	73
4.3.4 Modelo tres (Mal clasificado)	73
4.4.1 Censo y revisión de la información (control de calidad).	75
4.4.2 Análisis exploratorio de datos	75
4.4.2.1 Análisis por electrofacie	75
4.4.2.2 Análisis por modelo y por modelo-electrofacie	77
4.4.2.3 Conclusiones del análisis exploratorio de datos.	79
4.4.3 Análisis de agrupamiento	79
4.4.3.1 Conclusión del análisis de agrupamiento	87
4.4.4 Análisis de componentes principales y selección de variables.	89
4.4.4.1 Conclusiones del análisis de componentes principales.	95
4.4.5 Redes neuronales	95
4.4.5.1 Redes perceptrón multicapa. (MPL)	95
4.4.5.2 Resultados de la clasificación usando la red perceptron multicapa	101
4.4.5.4 Mapas auto-organizados.	103
4.4.5.4 Resultados de la clasificación con Mapas Auto-organizados (SOM)	107
4.4.5.5 Máquinas Comité (CM)	109
4.4.5.6 Conclusiones del entrenamiento con redes neuronales.	109
V. CASO DE ESTUDIO	111
5.1 Antecedentes	111
5.2 Información disponible.	112
5.3 Clasificación Manual	114
5.4 Aplicación de la metodología	115

5.4.1 Censo y revisión de la información (control de calidad).	115
5.4.2 Análisis exploratorio de datos	116
5.4.2.1 Análisis por electrofacie	116
5.4.2.2 Análisis de gráficos de caja (boxplots) por pozo	119
5.4.2.3 Análisis de gráficos de dispersión (gráficos de correlación o scatter plots)	119
5.4.2.4 Conclusiones del análisis exploratorio de datos.	119
5.4.3 Análisis de clústeres.	120
5.4.3.1 Conclusión del análisis de clústeres	127
5.4.4 Análisis de componentes principales y selección de variables.	127
5.4.4.1 Conclusiones del análisis de componentes principales.	129
5.4.5 Redes neuronales	130
5.4.5.1 Redes perceptrón multicapa.	131
5.4.5.2 Mapas auto-organizados.	136
5.4.5.3 Máquinas Comité.	141
5.4.5.4 Conclusiones del entrenamiento con redes neuronales.	144
5.4 Interpretación y comentarios de los resultados con redes neuronales	145
VI. ANÁLISIS Y DISCUSIÓN DE RESULTADOS	147
VII. PROPUESTA METODOLÓGICA.	150
7.1 Establecer un marco geológico de referencia.	150
7.2 Establecer los criterios de clasificación.	150
7.3 Control de calidad de los datos.	150
7.4 Análisis exploratorio de datos	150
7.5 Análisis agrupamiento	151
7.6 Análisis de componentes principales	152
7.7 Aplicación de las redes neuronales.	152
7.8 Metodología	154

CONCLUSIONES Y TRABAJOS FUTUROS	155
LISTA DE ABREVIATURAS	159
APÉNDICE A. ANÁLISIS EXPLORATORIO DE DATOS (AED) - REGISTROS SINTÉTICOS	160
APÉNDICE B. ANÁLISIS EXPLORATORIO DE DATOS (AED) - REGISTROS REALES	195
REFERENCIAS	218

ÍNDICE DE FIGURAS

Fig. 1.1 Flujo de trabajo de la Modelación Geológica Petrofísica de Yacimientos	5
Fig. 2.1. Neurona Biológica. Editada de Alpana 2000.	12
Fig. 2.2. Modelo de Neurona Artificial, Editada de Haykin, 1992	13
Fig. 2.3. Función Hardlimit simétrica	14
Fig. 2.4. Función Lineal.	14
Fig. 2.5. Función Sigmoidea.	15
Fig. 2.6. Tangente Hiperbólica.	15
Fig. 2.7. Rampa.	16
Fig. 2.8 Tipo de conexiones en una red monocapa	18
Fig. 2.9 Red Neuronal con tres capas ocultas y conexiones de alimentación hacia adelante.	18
Fig. 2.10 Red multicapa con diferentes tipos de conexiones	19
Fig. 2.11 Fronteras de decisión de las funciones OR y AND (tomado de Dunne, 2007).	24
Fig. 2.12 Superficie de error esquemática para una red perceptrón multicapa (tomado de Mejía, 2004)	26
Fig. 2.13 Arquitectura de la red perceptrón multicapa (tomado de Alpana, 2002).	27
Fig. 2.14 Trayectoria en el espacio de los pesos sinápticos	28
Fig. 2.15. Maldición de la dimensionalidad (Modificado de Bishop, 1995)	29
Fig. 2.16 Error y sobre entrenamiento.	30
Fig. 2.16 Arquitectura rectangular de un mapa auto-organizado	32
Fig. 3.1 Ejemplo de Histogramas normalizados por un petrofísico experto.	41
Fig. 3.2 Flujo trabajo propuesto	41
Fig. 3.3 Tipología de las distribuciones de frecuencias agrupadas	45
Fig. 3.4 Gráficos de cuantiles correspondientes a distintos tipos de distribuciones	45
Fig. 3.5 Transformación de las variables originales en factores o componentes principales	53
Fig. 3.6. Gráfico de sedimentación	55
Fig. 3.7 Esquema red neuronal perceptrón multicapa	61
Fig. 3.8. Esquema red neuronal de conjunto.	61
Fig. 3.9 Esquema de un módulo	62

Fig. 3.10 Esquema de la Máquina Comité	63
Fig. 3.11. Ejemplos de: a. Graficas de performance, b. histograma del error absoluto y c. matriz de confusión	64
Fig. 3.12 Ejemplos Gráfico de éxitos	64
Fig. 3.13 Ejemplos Mapa de distancia entre vecinos, Manual de Matlab (2009)	65
Fig. 3.14 Ejemplos de mapas de planos, para cuatro entradas, Manual de Matlab (2009)	65
Fig. 4.1 Patrones de apilamiento en la respuesta de los registros de rayos gama (electrofacies)	68
Fig. 4.2 Respuestas típicas de los registros convencionales de pozos, en litologías y tipos de fluidos específicos	68
Fig. 4.3 Modelo uno	72
Fig. 4.4 Registros geofísicos de Pozo sintéticos del Modelo uno.	72
Fig. 4.5 Modelo dos	73
Fig. 4.6 Registros geofísicos de Pozo sintéticos del Modelo dos.	74
Fig. 4.7 Modelo tres	74
Fig. 4.8 Registros geofísicos de Pozo sintéticos del Modelo tres.	75
Fig. 4.9 Dendrograma de los modelos uno, dos y tres, para una distancia menor a 40	80
Fig. 4.10 Dendrograma de los modelos uno, dos y tres, para una distancia menor a 40	80
Fig. 4.11 Dendrograma del modelo uno.	81
Fig. 4.12. Dendrograma del modelo dos	81
Fig. 4.13. Dendrograma del modelo tres	82
Fig. 4.14. Dendrograma del modelo uno. Utilizando los registros la relación RT/GR, Caliper, Neutrón y sónico	82
Fig. 4.15. Dendrograma del modelo dos. Utilizando los registros la relación RT/GR, Caliper, Neutrón y sónico	83
Fig. 4.16. Dendrograma del modelo tres. Utilizando los registros la relación RT/GR, Caliper, Neutrón y sónico	83
Fig. 4.17. Resultado de la clasificación por el método de k medias.	84
Fig. 4.18. Gráfico de puntos modelo uno, resultado de la clasificación por el método de k medias.	85
Fig. 4.19. Gráfico de puntos modelo dos, resultado de la clasificación por el método de k medias.	86

Fig. 4.20. Gráfico de puntos modelo tres, resultado de la clasificación por el método de k medias.	87
Fig. 4.21. Gráficos de dispersión, para la relación RT/GR y el registro sónico.	88
Fig. 4.22. Gráficos de dispersión, para la relación RT/GR y el registro neutrón.	89
Fig. 4.23. Gráficos de sedimentación	93
Fig. 4.24. Biplots	94
Fig. 4.25. Modelo de red neuronal	96
Fig. 4.26. Gráfico del monitoreo del error cuadrático medio	96
Fig. 4.27. Cuantificación del coeficiente de Pearson y gráfico de regresión lineal	97
Fig. 4.28. Matriz de confusión, de la estimación del modelo uno, entrenando con el modelo uno	97
Fig. 4.29. Gráfico del monitoreo del error cuadrático medio	98
Fig. 4.30. Cuantificación del coeficiente de Pearson y gráfico de regresión lineal	98
Fig. 4.31. Matriz de confusión, de la estimación del modelo dos, entrenando con el modelo uno	99
Fig. 4.32. Gráfico del monitoreo del error cuadrático medio	100
Fig. 4.33. Cuantificación del coeficiente de Pearson y gráfico de regresión lineal	100
Fig. 4.34. Matriz de confusión, de la estimación del modelo3, entrenando con el modelo1	101
Fig. 4.35 Resultados de la clasificación usando MPL	102
Fig. 4.36 Resultados de la clasificación, reinterpretando las salidas de la red MPL	102
Fig. 4.37 Mapa de pesos para el modelo uno	103
Fig. 4.38 Mapas de pesos para el modelo uno, por variable.	103
Fig. 4.39 Interpretación Mapa de éxitos del modelo uno	104
Fig. 4.40 Mapa de pesos para el modelo dos	105
Fig. 4.41 Mapas de pesos para el modelo dos, por variable	105
Fig. 4.42. Mapa de éxitos del modelo dos	106
Fig. 4.43 Mapa de pesos para el modelo tres	106
Fig. 4.44. Mapas de pesos para el modelo tres, por variable	107
Fig. 4.45. Interpretación Mapa de éxitos del modelo tres	107

Fig. 4.46. Resultados de la clasificación usando mapas auto-organizados	108
Fig. 4.47. Interpretación de la clasificación usando mapas auto-organizados	108
Fig. 4.48. Interpretación de la clasificación usando máquinas comité.	110
Fig. 5.1. Columna geológica del paleocanal de Chicontepec	112
Fig. 5.2. Distribución de los pozos	113
Fig. 5.3. Sección de los pozos utilizados	114
Fig. 5.4. Facies determinadas por especialistas, en el intervalo de inyección de CO ₂	115
Fig. 5.5 Dendrograma para todos los pozos	120
Fig. 5.6. Dendrograma de pozos 311, 312, 318 y 331.	121
Fig. 5.7. Dendrograma de pozos 332, 333, 351 y 353	121
Fig. 5.8. Resultado de la clasificación por el método de k medias, de los pozos 311, 312, 318 y 331	123
Fig. 5.9. Resultado de la clasificación por el método de k medias, de los pozos 332, 333, 351 y 353	123
Fig. 5.10. Gráfico de puntos modelo uno, resultado de la clasificación por el método de k medias.	124
Fig. 5.11. Gráficos de dispersión, para todos los pozos de los registros GR, RT, CALI, DT Y NPHI	125
Fig. 5.12. Gráficos de dispersión, para todos los pozos de los registros GR, RT, CALI, DT Y NPHI	126
Fig. 5.13. Gráficos de sedimentación, para los análisis de modelos en conjunto y para los modelos 1, 2 y 3 de manera individual.	128
Fig. 5.14. Biplot para el análisis de componentes principales y clústeres, usando todo los pozos.	129
Fig. 5.15. Biplots para el análisis de componentes principales y clústeres, para cada uno de los pozos	130
Fig. 5.16. Modelo de red neuronal	131
Fig. 5.17. Gráfico del monitoreo del error cuadrático medio.	132
Fig. 5.18. Cuantificación del coeficiente de Pearson y gráfico de regresión lineal	132
Fig. 5.19. Matrices de confusión, de la clasificación de los pozos 311, 312, 331 y 332.	133
Fig. 5.20. Matrices de confusión, de la clasificación de los pozos 333, 351 y 353	134
Fig. 5.21. Resultados de la clasificación usando MLP, para los pozos 311, 312, 318 y 331	135
Fig. 5.22. Resultados de la clasificación usando MLP, para los pozos 332, 333, 351 y 353	135
Fig. 5.23. Mapa de pesos del pozo 333	136

Fig. 5.24. Mapa de pesos del pozo 333, por variable	137
Fig. 5.25. Interpretación del mapa de éxitos del pozo 333	137
Fig. 5.26. Matrices de confusión, de la clasificación de los pozos 311, 312, 331 y 332.	138
Fig. 5.27. Matrices de confusión, de la clasificación de los pozos 333, 351 y 353.	139
Fig. 5.28. Resultados de la clasificación usando SOM, para los pozos 311, 312, 318 y 331	140
Fig. 5.29. Resultados de la clasificación usando SOM, para los pozos 333, 351 y 353	140
Fig. 5.30. Matrices de confusión, de la clasificación de los pozos 311, 312, 331 y 332.	141
Fig. 5.31. Matrices de confusión, de la clasificación de los pozos 311, 312, 331 y 332.	142
Fig. 5.32. Resultados de la clasificación usando CM, para los pozos 311, 312, 331 y 332	143
Fig. 5.33. Resultados de la clasificación usando CM, para los pozos 333, 351 y 353	144
Fig. 5.34. Sección de correlación de secuencias utilizando la clasificación de redes	146
Fig. A.1. Histogramas clase uno	162
Fig. A.2. Histogramas clase dos	163
Fig. A.3. Histogramas clase tres	164
Fig. A.4. Gráficos cuantil- cuantil clase	165
Fig. A.5. Gráficos cuantil- cuantil clase dos	166
Fig. A.6. Gráficos cuantil- cuantil clase tres	167
Fig. A.7. Gráficos de cajas por clase	168
Fig. A.8. Gráficos de dispersión de los modelos uno, dos y tres	169
Fig. A.9. Gráficos de caja por modelo	170
Fig. A.10. Histogramas modelo uno	173
Fig. A.11. Histogramas del modelo uno, clase uno	174
Fig. A.12. Histogramas del modelo uno, clase dos	175
Fig. A.13. Histogramas del modelo uno, clase tres	176
Fig. A.14. Gráfico cuantil- cuantil, modelo uno	177
Fig. A.15. Gráficos De dispersión Modelo uno	178

Fig. A.16. Histogramas del modelo dos	181
Fig. A.17. Histogramas del modelo dos, clase uno	182
Fig. A.18. Histogramas del modelo dos, clase dos	183
Fig. A.19. Histogramas del modelo dos, clase tres	184
Fig. A.20. Gráficos cuantil-cuantil, modelo dos	185
Fig. A.21. Gráfico de dispersión modelo dos	186
Fig. A.22 Histogramas del modelo tres	189
Fig. A.23. Histogramas del modelo tres, clase uno	190
Fig. A.24. Histogramas del modelo tres, clase dos	191
Fig. A.25. Histogramas del modelo tres, clase tres	192
Fig. A.26. Gráficos cuantil- cuantil, modelo tres	193
Fig. A.27. Gráfico de dispersión modelo tres	194
Fig. B.1 Histogramas litofacie uno(Areniscas)	197
Fig. B.2 Histogramas litofacie dos (Arenisca arcillosa)	198
Fig. B.3 Histogramas litofacie tres (Lutita – Arenosa)	199
Fig. B.4 Histogramas litofacie cuatro (Lutita)	200
Fig. B.5. Gráfico de cajas por litofacies	201
Fig. B.6 Gráficos cuantil- cuantil. litofacie uno (Arenisca)	202
Fig. B.7 Gráficos cuantil- cuantil. litofacie dos (Arena Arcillosa)	203
Fig. B.8 Gráficos cuantil- cuantil. litofacie tres (Lutita Arenosa)	204
Fig. B.9 Gráficos cuantil- cuantil. litofacie cuatro (Lutita)	205
Fig. B.10 gráficos de dispersión para todos los pozos.	206
Fig. B.11. Gráficos de caja por pozo	207
Fig. B.12 Gráfico de dispersión - Pozo 311	208
Fig. B.13. Gráfico de dispersión - Pozo 312	209
Fig. B.14. Gráfico de dispersión - Pozo 318	210

Fig. B.15. Gráfico de dispersión - Pozo 331	211
Fig. B.16. Gráfico de dispersión - Pozo 332	212
Fig. B.17. Gráfico de dispersión - Pozo 333	213
Fig. B.18. Gráfico de dispersión - Pozo 351	214
Fig. B.19. Gráfico de dispersión - Pozo 353	215

ÍNDICE TABLAS

Tabla 2.1. Clasificación de algunos modelos neuronales, en función del tipo de asociación de los datos de entrada y de salida.	22
Tabla 2.2 Funciones OR, AND y XOR, para dos variables lógicas M y N.	24
Tabla 4.1. Relación de las variables con cada una de las componentes, para los modelos uno, dos y tres	90
Tabla 4.2. Valores propios para cada componente, para los modelos uno, dos y tres	90
Tabla 4.3. Relación de las variables con cada una de las componentes, para los modelos 1	90
Tabla 4.4. Relación de las variables con cada una de las componentes, para el modelo uno	91
Tabla 4.5. Valores propios para cada componente, para el modelo uno	91
Tabla 4.6. Relación de las variables con cada una de las componentes, para el modelo uno	91
Tabla 4.7. Relación de las variables con cada una de las componentes, para el modelo dos	91
Tabla 4.8. Valores propios para cada componente, para el modelo dos	92
Tabla 4.9. Relación de las variables con cada una de las componentes, para el modelo dos	92
Tabla 4.10. Relación de las variables con cada una de las componentes, para el modelos 3	92
Tabla 4.11. Valores propios para cada componente, para el modelo tres	92
Tabla 4.12. Relación de las variables con cada una de las componentes, para el modelo tres	92
Tabla 5.1 Información disponible.	113
Tabla 5.2. Relación de las variables con cada una de las componentes, para los modelos uno, dos y tres	127
Tabla 5.3. Valores propios para cada componente	127
Tabla 5.4. Relación de las variables con cada una de las componentes	128
Tabla A.1 Estadísticos Clase uno	160

Tabla A.2 Estadísticos Clase dos	160
Tabla A.3 Estadísticos Clase tres	161
Tabla A.4. Estadísticos generales - Modelo uno	171
Tabla A.5. Estadísticos Modelo uno – Clase uno	171
Tabla A.6. Estadísticos Modelo uno – Clase dos	172
Tabla A.7. Estadísticos Modelo uno – Clase tres	172
Tabla A.8 Estadísticos Modelo dos	179
Tabla A.9 Estadísticos Modelo dos - Clase uno	179
Tabla A.10. Estadísticos Modelo dos – Clase dos	180
Tabla A.11. Estadísticos Modelo dos – Clase tres	180
Tabla A.12. Estadísticos Modelo tres	187
Tabla A.13. Estadísticos Modelo tres – Clase uno	187
Tabla A.14. Estadísticos Modelo tres – Clase dos	188
Tabla A.15. Estadísticos Modelo tres – Clase tres	188
Tabla B.1 Estadísticos litofacie uno (Arenisca)	195
Tabla B.2 Estadísticos litofacie dos (Arenisca arcillosa)	195
Tabla B.3 Estadísticos litofacie tres (Lutita arenosa)	196
Tabla B.4 Estadísticos litofacie cuatro (Lutita)	196

ÍNDICE DE CUADROS SINÓPTICOS

Cuadro sinóptico 3.1 Clasificación de algunos algoritmos jerárquicos	50
Cuadro sinóptico 3.2. Clasificación de algunos algoritmos no jerárquicos para el análisis de clústeres	51

INTRODUCCIÓN

Contar con un modelo adecuado de facies sedimentarias o cualquier modelo de clasificación de la información geológica-geofísica disponible, que nos permita diferenciar los diferentes eventos o cuerpos de rocas sedimentarias que se encuentran en un área prospectiva o de desarrollo de campos, tiene un impacto directo en muchas de las metodologías y en los flujos de trabajo, en las diferentes fases de la cadena de valor tanto de la exploración, como de la producción de hidrocarburos, desde definir un play hipotético y proponer un prospecto hasta la caracterización estática (para definir reservas) y dinámica de yacimientos (para proponer un esquema de desarrollo de campos). El modelado de facies proporciona el sustento del modelo sedimentario conceptual, el cual permitirá entre otras cosas reducir los riesgos e incertidumbres durante el modelado de cuencas, evaluar de los elementos del sistema petrolero de cualquier prospecto, así como comprender el comportamiento del flujo de fluidos dentro del espacio poroso durante la caracterización de un yacimiento.

Los procesos sedimentarios que definen las facies sedimentarias o depositacionales, gobiernan muchas de las características físicas y químicas de las rocas de interés dentro de la búsqueda de hidrocarburos. Definir el tipo de facies sedimentaria a la cual pertenece una roca dada, es un proceso de clasificación en sí mismo, cuyo criterio se basa en identificar rasgos o características de la roca que permiten definir un origen genético común (procesos físicos, químicos y biológicos durante el depósito de una roca). Generalmente estos rasgos como el contenido mineralógico, la clasificación textural, el contenido orgánico, las estructuras sedimentarias de diferentes órdenes, su color, etc., pueden llegar a ser visibles dependiendo de la escala en que sean analizadas, ya sea a escala de afloramiento, de muestras de canal, cortes de núcleo, láminas delgadas, etc. Para caracterizar el subsuelo es muy importante contar con núcleos, no obstante esta información no es continua (en algunos casos ni siquiera está disponible) y en el caso de las muestra de canal no es muy precisa (ya que depende de factores operativos durante la perforación); por lo que se requiere contar con información continua. Lo anterior ha llevado a utilizar información indirecta de las rocas utilizando herramientas como los registros geofísicos de pozo, información sísmica u otro método de prospección geofísica, las cuales responden a propiedades físicas de la roca, que muchas veces (más no siempre) pueden ligarse a eventos geológicos que caracterizan el proceso depositacional de las rocas.

Es importante resaltar que las mediciones indirectas que ofrecen los métodos geofísicos ya sean superficiales o de pozo, permiten establecer criterios de clasificación de las rocas que han sido medidas, en función de las propiedades físicas que representan y/o de la escala de medida (el ejemplo más claro es la diferencia entre el registro sísmico y la información sísmica, las cuales explotan la misma propiedad física). No obstante estas propiedades físicas no necesariamente pueden ser interpretadas y clasificadas de forma directa como facies de depósito, ya que muchos efectos físicos externos influyen en estas respuestas (compactación, presión, temperatura). Muchas veces son utilizados otros conceptos para realizar clasificaciones en función de la litología (litofacies), en el contenido biológico (biofacies), en la respuesta sísmica (sismofacies), etc. En el caso de los registros de pozo el término comúnmente utilizado es el de electrofacies y es el enfoque utilizado en este trabajo de tesis. El definir electrofacies es un proceso de clasificación, basado en criterios definidos a partir de registros de pozos (utilizando registros que respondan a litologías, a fluidos, propiedades petrofísicas como la permeabilidad o porosidad y/o patrones de apilamiento, siendo este último el más utilizado para inferir sistemas o ambientes de depósito), el cual está en función del objetivo que se persiga, no sin olvidar antes que la clave del éxito dentro de diferentes procesos exploratorios y de caracterización de un yacimiento, es la integración de toda la información disponible.

El criterio de clasificación de registros geofísicos de pozos utilizado en este trabajo de tesis, es principalmente el litológico (clasificación de los registros que respondan principalmente a la litología), ya que el enfoque que se utiliza es el de la metodología de la caracterización estática o también conocida como Modelación Geológica Petrofísica de Yacimientos (MGPY).

El objetivo principal de este trabajo de tesis, es probar una metodología de análisis estadístico, que permite dar un mejor conocimiento acerca de los datos a clasificar, independientemente de la metodología de clasificación (manual o automática). El trabajo de clasificación manual es un proceso

largo, detallado y demandante, que depende de los criterios del analista, por lo que dicho proceso se vuelve impráctico cuando aumenta el número de pozos a analizar, por lo que el uso de herramientas computacionales automatizadas ha sido muy popular. Diferentes metodologías computacionales de clasificación de datos son implementadas en algunas aplicaciones comerciales, sin embargo, un uso adecuado de estas técnicas depende de la calidad de los datos de entrada y su naturaleza, por lo que en este trabajo se propone como punto medular, el uso del análisis exploratorio de datos y diferentes técnicas estadísticas para plantear un proceso de clasificación adecuado.

Para ilustrar los alcances del análisis estadístico de este trabajo de tesis, se propone el uso de las redes neuronales como algoritmos clasificadores de electrofacies, Existen diversos e innumerables trabajos para resolver el problema de clasificación de electrofacies mediante el uso de las redes neuronales. En esta tesis es propuesto el uso de tres modelos neuronales distintos, las llamadas redes perceptrón multicapa, los mapas auto-organizados o redes de Kohonen, llamadas así en honor del profesor Teuvo Kohonen; además es planteado un esquema de entrenamiento de grupos o módulos de redes neuronales, propuesto en Alpana (2000), las cuales en conjunto resuelven problemas de clasificación; este esquema de entrenamiento, es propiamente un modelo neuronal y se le conoce con el nombre las maquina comité (committee machine).

En el primer capítulo de este trabajo se establece la importancia de la clasificación de facies y el uso de las redes neuronales en la clasificación de facies, en el contexto de la Modelación Geológica Petrofísica de Yacimientos (MGPY).

En el segundo capítulo, son establecidos los antecedentes teóricos generales de las redes neuronales y los diferentes modelos propuestos. Se bosquejan también los antecedentes de la clasificación de litofacies, mediante técnicas computacionales y el uso de las redes neuronales en la clasificación de litofacies.

En el tercer capítulo, son mostradas y descritas (tanto los aspectos teóricos de las técnicas utilizadas, como los aspectos prácticos de las mismas) las etapas de la metodología propuesta.

En el cuarto capítulo, se aplica y analiza la metodología propuesta a datos de registros geofísicos de pozo sintéticos, con el fin de mostrar el proceso de clasificación mediante la metodología aplicada a tres casos controlados con niveles de clasificación distintos (datos bien, regular y mal clasificados).

En el quinto capítulo la metodología es aplicada a datos de pozos reales en un campo dentro de la cuenca de Chicontepec. En el capítulo seis los análisis de los resultados obtenidos son discutido y partir de estos análisis, es planteada la propuesta metodológica de mostrada en el capítulo siete.

es planteada una propuesta metodológica final, fundamentada en el análisis de los resultados obtenidos en los capítulos cuatro y cinco.

Este trabajo pretende mostrar la utilidad y ventaja de los análisis estadísticos previos a un proceso de clasificación, pero además pretende explotar las bondades que brindan las redes neuronales como sistemas de clasificadores optimizando su aplicación mediante el enfoque modular de las “*máquinas comité*”, obteniendo mejor grado de clasificación en los casos que los mismos datos lo permitan. El uso de las redes neuronales nos permite dar consistencia a las clasificaciones propuestas de manera manual y estimar con cierto grado de aproximación las electrofacies en pozos que no hayan sido clasificados.

I. LAS REDES NEURONALES ARTIFICIALES EN EL MARCO DE LA MODELACIÓN GEOLÓGICA PETROFÍSICA DE YACIMIENTOS (*MGPY*)

La modelación geológica petrofísica (*MGPY*) o también conocida como caracterización estática, es una tarea multidisciplinaria encargada de integrar la información disponible de un yacimiento en un modelo numérico que represente las propiedades petrofísicas en función de su geología (Casar et. Al., 2012). Uno de los principales supuestos durante el proceso de caracterización de un yacimiento, establece que los datos conocidos nunca son suficientes, ni exhaustivos para describir en su totalidad un yacimiento de manera completa y exacta (Casar et. al., 2012), ya que la información de pozos suele estar limitada horizontalmente (nunca se tiene un muestreo continuo de pozos) y la información sísmica suele tener limitantes de resolución tanto vertical como horizontal. Bajo este enfoque, en diversos trabajos (Cosentino, 2001; Deutsch, 2002; Kelkar y Pérez, 2002; Caers, 2005), se puede establecer una metodología estándar para la *MGPY*, la cual hace uso de herramientas matemáticas – computacionales (principalmente estocásticas), que permiten estimar la distribución del tipo de rocas y sus propiedades petrofísicas dentro de un yacimiento.

1.1 Modelación Geológica Petrofísica de Yacimientos (*MGPY*)

La *MGPY* es un proceso integral (Cosentino, 2001), ya que tiene como finalidad combinar información de diversas fuentes de manera integrada, desde el punto de vista conceptual y numérico. Es debido a lo anterior que el uso de técnicas geoestadísticas facilita esta integración y permiten analizar la influencia de los diversos parámetros que intervienen en el modelo, obtener un modelo numérico confiable y realista de los yacimientos, permitiendo realizar un análisis del grado de incertidumbre asociado a los modelos obtenidos (Casar et. Al., 2012).

Es importante señalar que la *MGPY* parte de un modelo geológico conceptual, el cual incluye el ambiente geológico depositacional y un marco estratigráfico producto de la sedimentación y formación de las rocas que constituyen el yacimiento. El modelo geológico conceptual debe incluir la correlación de las unidades estratigráficas lo cual involucra disciplinas tales como: interpretación sísmica, estratigrafía de secuencias, sedimentología, interpretación de registros de pozos, bioestratigrafía, y estudios geológicos de análogos de afloramientos. El modelo geológico conceptual del yacimiento debe proporcionar una evaluación cuantitativa de los parámetros tales como geometría, extensión y distribución de la facies y la relación que guardan entre sí (Casar et. Al., 2012).

En la parte estructural es necesario tener identificadas y definidas las características estructurales del yacimiento, como son las fallas regionales, fallas locales y fracturas. La información petrofísica es obtenida de las curvas de los registros geofísicos de pozo, por lo general calibrados con la información petrofísica de los núcleos, la cual es obtenida directamente de la roca en pruebas de laboratorio. La integración de datos dinámicos como son pruebas de presión, pruebas de trazadores y datos de producción, ofrecen información relativa al flujo a gran escala y debe de ser tomada en cuenta para en lo posible incluirla en los modelos geológicos conceptuales (Casar et. Al., 2012).

1.2 Componentes de la *MGPY*

La modelación geológica-petrofísica consta de dos partes: el modelo geológico y el modelo petrofísico (Casar et. Al., 2012)

a) Modelo geológico

Refleja el modelo geológico conceptual del yacimiento, el cual consta de un marco estructural y un modelo de depósito, el cual debe ser consistente con el modelado geológico regional.

Dentro de la *MGPY*, el modelado geológico consta de una geometría y una arquitectura (modelo estructural), una malla estratigráfica (modelo estratigráfico), un modelo de distribución de facies (modelo de depósito), así como la caracterización de las heterogeneidades del yacimiento.

b) Modelo Petrofísico

Plasma la conceptualización de la roca a partir de su comportamiento físico derivado de las propiedades petrológico-sedimentarias de la roca (tamaño de grano, clasificación textural, mineralogía, cementación, diagénesis, contenido arcilloso, etc.), el contenido de fluidos y las condiciones propias del yacimiento.

Principalmente el modelo petrofísico establece el vínculo existente entre la red de poros, su contenido de fluidos y el flujo, como la parte medular del estudio de los yacimientos, ya que las propiedades petrofísicas son las que en última instancia controlan el flujo y transporte de fluidos en el yacimiento (Casar et. al., 2012). Las principales propiedades petrofísicas consideradas dentro de la *MGPY* son: la porosidad, la saturación de agua, permeabilidad y volumen de arcilla.

1.3 Flujo de trabajo de la *MGPY*

El flujo de trabajo de la modelación geológica petrofísica, considerado en este trabajo es el presentado en Casar et. al. (2012), el cual consta de dos etapas (Fig. 1.1):

A. Conocimiento geológico

Esta etapa es previa al *MGPY*, en esta etapa se recaba y se analiza información geológica y se interpreta información proveniente de métodos de exploración. Se definen facies y se crea un modelo geológico conceptual.

- **Geología:** Conocimiento geológico de área o yacimiento, estudio de análogos de yacimiento en afloramiento, definición del sistema de depósito, definición de facies, descripción y estudio de núcleos, definición del sistema poroso, análisis de registros geofísicos de pozo, identificación de facies en registros de pozo, determinación de propiedades petrofísicas en registros de pozo, determinación de cima y base de las principales unidades que forman el yacimiento en registros de pozo. Definición y dimensiones de la geometría de los cuerpos sedimentarios, secciones de correlación de cuerpos sedimentarios y facies, espesor promedio de las unidades estratigráficas. Identificación de los procesos diagenéticos y sus implicaciones. Estudio de patrones de fracturas y medición de parámetros (orientación, apertura, densidad, etc.).
- **Sísmica:** Conocimiento del modelo estructural de yacimiento, características estructurales, definición de fallas regionales, fallas locales. Definición de las superficies que delimitan el yacimiento y de posibles unidades de flujo. Distribución lateral de los cuerpos sedimentarios. Derivación de atributos sísmicos para correlación con propiedades petrofísicas.
- **Heterogeneidades:** Conocimiento sobre las heterogeneidades del yacimiento con posible impacto en el flujo de los hidrocarburos. Definición de patrones de fracturas. Identificación del medio poroso.
- **Datos dinámicos:** Conocimiento e interpretación de datos dinámicos como son pruebas de presión, trazadores, datos de producción y pruebas piloto de inyección.

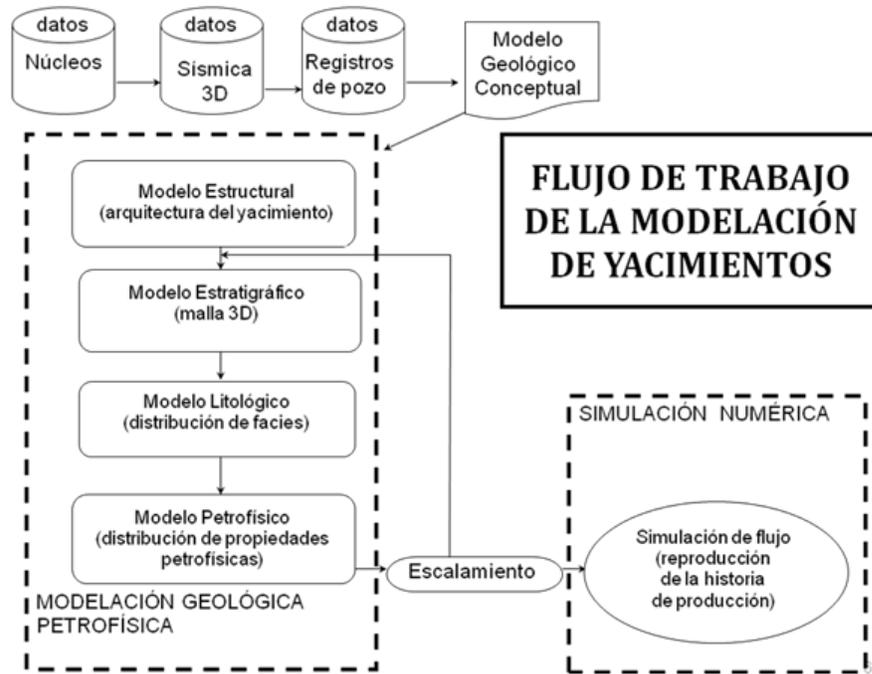


Fig. 1.1 Flujo de trabajo de la Modelación Geológica Petrofísica de Yacimientos

B. Modelo Geológico-Petrofísico

Esta es la etapa propiamente de la *MGPY*, consta de un análisis exploratorio de datos, geometría y arquitectura del yacimiento, construcción de la malla estratigráfica, análisis heterogeneidades, modelo de distribución de facies, modelo de distribución de propiedades petrofísicas y análisis de incertidumbre.

- **Análisis Exploratorio de Datos:** Análisis estadístico de la información de facies y propiedades petrofísicas, identificación de facies estadísticamente comparables para fines de agrupamiento en litotipos. Creación de facies en dos categorías: productora, no productora. Tratamiento estadístico multivariado de los datos para formar agrupamientos. Detección de valores atípicos. Transformación de variables. Análisis de simetría y análisis de la distribución de probabilidad de la información. Identificación de tendencias. Análisis estadístico de correlación entre variables.
- **Geometría y arquitectura del yacimiento**

Consiste en establecer e identificar la arquitectura básica del yacimiento, es decir se deben de definir las fronteras del depósito. La arquitectura del yacimiento involucra identificar e interpretar las principales fallas, así como las superficies que representan la cima y la base del yacimiento. Más aún se deben identificar las fallas que pueden actuar como unidades de flujo o aquellas que constituyen barreras impermeables. Mientras que las características estructurales como las fallas regionales se definen de manera determinística usando datos sísmicos, las características de las fallas locales y/o fracturas pueden ser simuladas usando modelos estocásticos.

- **Construcción de la malla estratigráfica**

Esta parte se encarga de definir el marco o estructura interna del yacimiento, el proceso considera la identificación y definición de las superficies que delimitan a las principales unidades de flujo del yacimiento a partir de la correlación de unidades geológicas en los pozos. El modelo geológico conceptual del yacimiento provee de un medio adicional para inferir la longitud de correlación de las

facies y las dimensiones y espesores promedio de las unidades. Es necesario definir la resolución de la malla en el sentido vertical, así como la geometría interna de las unidades (capas continuas, paralelas truncadas, etc.); este proceso es de suma importancia ya que los registros se escalan a esta resolución, intentando no perder información durante este proceso.

- **Heterogeneidad del yacimiento**

Las heterogeneidades del yacimiento son características geológicas de escala menor, que no son resueltas por la resolución de la malla estratigráfica y por lo tanto no significativas desde el punto de vista estrictamente estático de la caracterización del yacimiento, pero tienen un impacto significativo en el flujo de los hidrocarburos; tal es el caso de las fracturas y los vórgulos en los yacimientos carbonatados. En el caso de yacimientos fracturados, los diferentes sistemas de fracturas se simulan mediante un modelo de redes de fracturas discretas, ya sea por facies y/o por unidades estratigráficas.

- **Modelado de la distribución de facies**

Las facies en el yacimiento representan la guía de la distribución de sus propiedades petrofísicas, ya que en la mayoría de los casos estos dos aspectos están altamente relacionados (Díaz, 2006; Casar et. al., 2012)).

El modelo de distribución de facies se basa en el modelo geológico conceptual y en el modelo estratigráfico-sedimentológico del yacimiento; sin embargo, el criterio de clasificación, depende de la finalidad que se tengan. Por ejemplo, el concepto de facies es particularmente adecuado para los estudios integrales de yacimientos, ya que pueden considerarse como el volumen elemental práctico del yacimiento, y representan el bloque básico para la construcción de modelos geológicos en 3-D, tomando en cuenta el aspecto litológico.

El término facie se ha definido por diversos autores, en diferentes tiempos y en la actualidad posee diversas concepciones; no obstante, un punto en común ha sido, que para caracterizar una facie, es necesario tener un conjunto de propiedades que sean comunes a una unidad estratigráfica, por lo que podemos hablar de litofacies, electrofacies, sismofacies, icnofacies, etc. En general el término facie se asocia a las características del depósito sedimentario que representan un origen genético común. Una facie depositacional es definida por Moore (1949) como cualquier parte arealmente restringida de una unidad estratigráfica, la cual exhibe características significativamente diferentes, de otras partes de la misma unidad estratigráfica.

Las facies de depósito se asocian principalmente a condiciones físico-químicas y biológicas, específicas durante el depósito del sedimento (ambiente de depósito). Tradicionalmente, las facies de depósito, se identifican por medio de la observación directa de núcleos o por las características de las rocas visualizadas en afloramientos, sin embargo, la carencia de núcleos en un área representativa del yacimiento limita su estudio (Gonzales y Mendez, 2007), por lo que ha sido necesario identificar rasgos medibles en forma indirecta. Una práctica común para identificar facies de depósito es la interpretación de patrones geométricos en los registros de rayos gama y de resistividad, para asociarlos a un sistema de depósito conceptualizado a partir de la integración de la información disponible en un área (sísmica, núcleos, muestras de canal, datos paleontológicos, etc).

La litología está ampliamente relacionada con el tipo de facie sedimentaria, no obstante, una facie de depósito puede constituirse por secuencias de litologías diferentes. En general para el tipo de litología se emplea el termino litofacie. Comúnmente las litologías son identificadas del estudio de láminas delgadas y núcleos, además de ser inferidas del análisis los registros rayos gama, resistividad, factor fotoeléctrico, densidad, neutrón y sónico. Los especialistas utilizan crossplots de los registros densidad-neutrón, densidad-sónico y sónico-neutrón, para construir complejos modelos petrofísicos multiminerale y así inferir las litologías.

Uno de los problemas principales del uso de registros de pozo, es que éstos se ven afectados por diversos factores como lo son el tipo de fluido, los efectos diagenéticos, la sobrepresión, las fallas, etc., por lo que comúnmente se utiliza el término electrofacies para definir un criterio de clasificación basado únicamente en los valores de los registros de pozo. El término electrofacie fue introducido por Serra y Abbott (1980) y ha sido definido como un conjunto de respuestas de los registros geofísicos de pozo, el cual caracteriza un estrato y permiten distinguirlo de otros. Las electrofacies, también llamadas facies de registros geofísicos, son respuestas de medidas de propiedades físicas de las rocas (Bishnu y Mahendra, 2006), es decir, una combinación de repuestas indirectas de litologías y fluidos.

Existen otros criterios de clasificación basados en las propiedades petrofísicas evaluadas e interpretadas (mineralogía, permeabilidad, volumen y tipo de arcillas, saturación de fluidos, etc), a partir de los registros de pozo, sin embargo este tipo de clasificaciones (como ejemplo de dichas clases podríamos tener: arenas compactas, arenas permeables, arenas arcillosas, arenas saturadas de agua, de hidrocarburo, arenas porosas, lutitas, carbonatos fracturados, etc.), son más utilizadas en problemas específicos, dependiendo del objetivo y del interprete.

En la *MGPY*, es necesario establecer un criterio de clasificación (facies de depósito, litofacies o en función de sus propiedades petrofísicas), el cual comúnmente se simplifica mediante la agrupación de litotipos y/o de clases petrofísicas, reduciendo de manera práctica la clasificación a dos tipos de facies: la almacenadora y la no almacenadora. Durante esta etapa se generan esquemas de clasificación de facies extendidos a todos los pozos. Secciones de correlación de facies. La estadística y curvas de probabilidad de ocurrencia de facies en sentido vertical. Se crean los mapas de probabilidad de ocurrencia de facies en sentido horizontal. Mapas de isopacas de las unidades estratigráficas o de ciclos sedimentarios para guiar la modelación. Modelo de correlación espacial en 3D para facies.

- **Simulación estocástica de facies**

Simulación de facies mediante métodos apropiados para facies, principalmente métodos geoestadísticos (Simulación Secuencial de Indicador, Simulación Gaussiana Truncada, etc.)

- **Modelo de distribución de propiedades petrofísicas:**

Definición de propiedades petrofísicas en pozos. Escalamiento de los registros petrofísicos a la resolución de la malla en sentido vertical. Análisis exploratorio de la información petrofísica escalada.

- **Análisis estadístico de la correlación entre propiedades petrofísicas.**

Análisis de correlación espacial para propiedades petrofísicas. Estimación y modelado de variogramas en dirección horizontal y vertical para cada variable.

- **Simulación estocástica de propiedades petrofísicas**

Simulación de propiedades petrofísicas mediante modelos Gaussianos.

- Análisis de incertidumbre: Evaluación de escenarios mediante múltiples realizaciones de las variables en estudio. Escenario optimista, percentil del 90% (P90), escenario medio, percentil del 50% (P50) y escenario pesimista, percentil del 10% (P10). Mapas de probabilidad de ocurrencia de variables de interés.

1.4 Redes neuronales en la *MGPY*

Las redes neuronales artificiales (RNA), son sistemas computacionales que pertenecen al ramo de la inteligencia artificial (IA). Esta tecnología son sistemas alternos a los métodos estadísticos tradicionales,

para la clasificación, optimización y estimación de funciones, sin embargo como se detallará en el capítulo siguiente, la naturaleza de las RNA es en sí misma estadística.

Existen numerosos trabajos asociados a la aplicación de las redes neuronales artificiales en el campo de las geociencias. Mucho del software comercial dentro ramo petrolero, ha hecho uso de esta tecnología con el fin de brindar soluciones a problemas relacionados con procesos de clasificación y estimación de datos (Stratimagic de Paradig®, Petrel de Schlumberger®, Hampson and Rusell de CGG®, entre otros), principalmente su uso ha sido para clasificación de registros de pozo y forma de onda sísmica, para estimar curvas en registros de pozo y reconocer patrones para identificar fallas en los datos sísmicos, también se han aplicado para procesos de solución de problemas inversos (optimización), en conjunto con otros algoritmos, para crear cubos de propiedades a partir de pozos (EMERGE - Hampson and Rusell, Genetic Inversion – Petrel, etc.)

La aplicación que podría dársele a las redes neuronales, dentro del proceso de la Modelación Geológica Petrofísica de Yacimientos es principalmente la de clasificación de facies (tanto sísmicas como electrofacies), estimación de curvas completas o en tramos faltantes de registros a partir de curvas de otros del mismo pozo y/u otros pozos, propagación de propiedades petrofísicas mediante el reconocimiento de patrones.

En lo que respecta a la estimación de facies (de depósito, litofacies, clases petrofísicas o cualquier otra clasificación), existen numerosas técnicas capaces de resolver el problema, con sus pros y sus contras. Los más utilizados en la actualidad son análisis estadísticos multivariados principalmente el análisis de agrupamiento o clústeres (AA), con sus diferentes algoritmos (métodos jerárquicos y no jerárquicos), así como el uso creciente de los sistemas neuronales. Sin embargo existen muchas más metodologías para la clasificación de datos, por ejemplo lógica difusa (Interactive Petrophysic – Senenergyworks, 2013) y algoritmos genéticos (Robert John y Birkenhead Ralph, 2001), además de muchos sistemas híbridos.

II. REVISIÓN DEL ESTADO DEL ARTE DE LA APLICACIÓN DE LAS REDES NEURONALES ARTIFICIALES A LA CLASIFICACIÓN DE FACIES

El Hombre a través de los tiempos ha pensado en la forma de resolver problemas de manera más rápida y eficiente (López, 2001). Motivado por el interés de comprender el funcionamiento y la capacidad que tiene del cerebro humano, en la solución de problemas complejos, junto con el creciente desarrollo de la tecnología computacional, se originaron los modelos conocidos como Redes Neuronales artificiales (Dunne, 2007).

Las redes neuronales artificiales, se encuentran dentro del área de estudios de las ciencias computacionales, conocido como Inteligencia Artificial. El estudio de las redes neuronales es un campo de investigación en el desarrollo de modelos matemáticos y computacionales a partir de elementos simples y no lineales (Corchado, et. al., 2000; Hilera y Martínez, 1995; Fausett, 1994).

Su aplicación se ha extendido a diversas disciplinas y ciencias como la informática, la electrónica, las telecomunicaciones, la economía, la medicina, la geología, la criminalística, la ingeniería ambiental, entre muchas otras.

Sin embargo de manera general, los tipos de problemas que han sido más estudiados son (Fyfe, 1996):

1. Clasificación y reconocimiento de Patrones
2. Optimización
3. Estimación y aproximación de funciones
4. Compresión de Datos.

Desde la perspectiva del reconocimiento de patrones, las redes neuronales pueden ser consideradas como una extensión de las muchas técnicas convencionales (patrones estadísticos) que han sido desarrolladas durante varias décadas (Bishop, 1995). A pesar de que el comportamiento de las redes neuronales deriva de una analogía biológica, el enfoque teórico más natural dentro del reconocimiento de patrones es el estadístico, el cual reconoce la naturaleza probabilística tanto de la información que buscamos procesar y de la forma en que se presentan los resultados, por lo que el uso que este campo augura poderosos resultados (Bishop, 1995).

2.1 Historia de las Redes Neuronales Artificiales (RNA).

El estudio del cerebro humano data de cientos de años atrás. Los primeros trabajos encaminados al desarrollo de las redes neuronales artificiales, se atribuyen a Warren McCulloch, un neurofisiólogo y a un joven matemático llamado Walter Pitts, en 1943. Ellos escribieron un artículo, explicando el funcionamiento de la neurona. Ellos modelaron una red neuronal simple, empleando circuitos eléctricos. En 1949 Donald Hebb, retomó el trabajo de McCulloch y Pitts y escribe su libro llamado "Organización del comportamiento", donde establece que la rutas neuronales de información, se fortalecen cada vez que son utilizadas.

Con el avance de los computadores en los años 50's se vuelve posible comenzar a modelar las teorías concernientes al pensamiento humano. Nathaniel Rochester de los laboratorios de investigación de IBM, dirigió los primeros intentos por simular una red neuronal, los cuales fueron fallidos en un inicio, pero progresivamente fueron teniendo éxito. Marvin Minsky y Dean Edmons, diseñaron las SNARC (Stochastic Neural Analog Reinforcement Calculator), mediante el uso de motores y dispositivos mecánicos, generando una máquina capaz de aprender.

En 1956 el investigador Dartmouth Summer dio un gran impulso tanto a la inteligencia artificial, como a las redes neuronales y sentó las bases de los trabajos de John Von NeumRNA. NeumRNA propuso imitar las funciones de las neuronas simples mediante el uso de tubos de vacío. En esos instantes Frank Rosenblatt, un neurobiólogo de la Universidad de Cornell, comenzó a trabajar en el perceptrón, inspirado

en la visión que tendría una mosca. El perceptrón de Rosenblatt fue construido únicamente con hardware, siendo esta la red neuronal más antigua y actualmente con muchas aplicaciones.

En 1959 Bernard Widrow y Marcian Hoff de Stanford, desarrollaron el modelo llamado ADALINE (Adaptative Linear Element) y MADALINE (Múltiple Adaptative Linear Element), las cuales fueron llamadas Elementos lineales adaptativos múltiples. El modelo MADALINE fue la primera red neuronal aplicada a problemas reales, la cual sigue teniendo aplicaciones comerciales.

En 1969 Marvin Minsky y Seymour Papert publicaron el libro "Perceptrones", en donde se comenzó a estudiar las limitaciones de los modelos neuronales artificiales. Minsky y Papert mostraron que las unidades de procesamiento como el perceptrón o el ADALINE, únicamente pueden ser aplicados a problemas que sean linealmente separables. El perceptrón es útil para clasificar un conjunto de entradas continuas en dos clases únicamente, ya que calcula una suma ponderada de las entradas, substraer un valor umbral y arroja dos valores posibles como salida, limitando así su aplicación.

Se le ha atribuido al libro "Perceptrones" el receso en la investigación de las redes neuronales ocurrido a lo largo de las décadas de los 70's y 80's. Desafortunadamente la exageración sobre el potencial de las redes neuronales, las limitaciones reales en la aplicación de los modelos existentes y los aspectos éticos relacionados a "máquinas pensantes", generaron que se dejarán de financiar proyectos asociados al desarrollo de las redes neuronales, periodo que perduró hasta 1981. Sin embargo algunos investigadores, sobre todo en la psicología cognitiva, continuaron con sus investigaciones e hicieron contribuciones importantes en la década de 1970.

James Anderson de la Universidad Brown, Stephen Grossberg la Universidad de Boston, Paul Werbos de la Universidad de Harvard, Teuvo Kohonen en el Instituto de Tecnología de Helsinki, John Hopfield del Instituto Tecnológico de California (CALTECH), David Rumelhart universidad de California y Geoffrey Hinton de la Universidad de Pittsburg, realizaron grandes avances en diferentes campos de aplicación.

A principios de los 70's James Anderson desarrolló un modelo lineal llamado asociador lineal y posteriormente en 1977 una poderosa extensión conocida como la red tipo BSB (BrainState in a Box). En 1974 Werbos sentó las bases para el desarrollo del algoritmo de retropropagación y ya en 1977, Stephen Grossberg, desarrolló la teoría de Resonancia Adaptada (TRA), donde la arquitectura de la red simula otras habilidades del cerebro: memoria a largo y corto plazo. Kohonen, en 1978, continuó su investigación sobre cómo los mapas topológicos sensoriales en el cerebro podrían ser recreados con una máquina, lo que lleva a las famosas redes de Kohonen y los mapas auto-organizativos. En 1982 Hopfield, quien con claridad matemática, mostró lo que las redes recurrentes llamadas autoasociativas (de memoria asociativa o simplemente red de Hopfield) podían hacer, despertando nuevamente el interés por la investigación sobre los modelos neuronales, a partir de la publicación de su libro "Computación neuronal de decisiones en problemas de optimización", en 1985. Rumelhart y Hinton, reformularon el algoritmo de aprendizaje de retropropagación, en 1982.

A partir de 1986, el panorama fue alentador con respecto a las investigaciones y el desarrollo de las redes neuronales. Se comenzaron a celebrar conferencias y congresos anuales sobre redes neuronales, organizadas por instituciones como el Instituto Americano de Física, el Instituto Americano de Ingenieros Eléctricos y Electrónicos, Asociaciones de Ingenieros Japoneses y Europeos, así como el Departamento de Defensa de los Estados Unidos, entre otros. En la actualidad, son numerosos los trabajos que se realizan y publican cada año, surgen nuevas aplicaciones y las empresas lanzan al mercado nuevos productos, tanto hardware como software (sobre todo para simulación).

2.1.1 Las RNA en la predicción de Litofacies

La identificación y zonificación de electrofacies, es un problema de clasificación y/u optimización; y como un requisito previo los registros continuos primero tiene que ser segmentado en zonas discretas de propiedades similares, que son las unidades elementales de referencia para inferir la correlación entre los pozos. El contexto y los objetivos del análisis determinan la naturaleza de las zonas, las cuales pueden

ser asignadas como unidades litoestratigráficas, petrofísicas u otras entidades geológicas significativas (Alpana, 2002).

Las redes neuronales se han convertido en una herramienta popular que facilita y apoya la clasificación de electrofacies (interpretadas como litofacies), sin embargo, es necesario entenderlas y entender la naturaleza propia de los datos, así como el aporte humano para la interpretación e integración de los resultados. En el presente trabajo de tesis, se pretende focalizar esta filosofía de trabajo, para la optimización de los resultados obtenidos.

Desde los primeros días de la introducción de las computadoras a la geología, los procedimientos numéricos computarizados, mezclados con el razonamiento de un especialista experto, han sido empleados para simular los criterios heurístico-intuitivos del reconocimiento de patrones para la zonificación. Debido a la naturaleza de este problema, los métodos usualmente más utilizados son: el análisis de componentes principales, el análisis de factores discriminantes, el análisis multivariado, las regresiones no paramétricas, el análisis de agrupamiento (clústeres) y/o técnicas de inteligencia artificial

Existen varias ventajas en estos procedimientos computarizados. Los métodos numéricos son más potentes que el cerebro humano, en la comprensión simultánea de un conjunto de datos multivariantes (Alpana, 2002). Por otra parte, los resultados llegan a ser objetivos, coherentes, reproducibles y libre de criterios personales realizados por el analista de registros.

En la actualidad existen numerosos trabajos, encaminados a utilizar diversas técnicas computacionales, que faciliten al analista el proceso de clasificación de electrofacies. Por mencionar algunos Wolff y Pelissier-Combescure (1982) utilizaron FACIOLOG para la zonificación de registros utilizando análisis de componentes principales y análisis de distribución modal.

En 1987, Delfiner y Busch utilizaron el análisis de factor discriminante para la correlación de los valores de registro con las unidades litológicas. En estos trabajos se hace hincapié en la selección de los parámetros que proporcionan la mayor discriminación.

Fang et al. (1992) empleó la técnica de reconocimiento de patrones, mediante correlación de onda dinámica, para identificar electrofacies, siendo éste un proceso supervisado, donde se compara un registro ya clasificado, con otro a clasificar. Gill et al. (1993) utilizó la técnica de análisis de agrupamiento, para llegar al mismo objetivo. Este método estadístico se basa en la discriminación de grupos (clústeres) de electrofacies de tal manera que dentro de cada grupo la dispersión o distancia de los datos es mínima.

Las redes neuronales en la clasificación de facies fueron introducidas durante los primeros años del desarrollo de la tecnología. Rogers et al. (1992) aplicó una red neuronal de alimentación hacia adelante (feedforward Neural Network), también llamadas redes de retropropagación (Backpropagation Artificial Neural Network-*BPRNA*), para clasificar la litología mediante registros de pozos. La configuración de la red fue de tres neuronas en la capa de entrada, que corresponde a los registros rayos gamma, densidad y porosidad de neutrón. La capa oculta constaba de tres y cuatro neuronas y la capa de salida consta de cuatro neuronas que corresponden a la clasificación de cuatro litofacies: caliza, dolomita, lutita y arenisca. En un estudio comparativo Wong et al. (1995), mostró mejores resultados en la clasificación de electrofacies, usando el análisis del factor discriminante, sobre los resultados obtenidos con el entrenamiento de una red de retropropagación Zhang et al. (1999) identificó litofacies mediante la integración de redes de retropropagación y redes neuronales llamadas mapas auto-organizados. Los datos del registro se dividieron en intervalos de acuerdo a las litofacies identificadas a partir de los núcleos y posteriormente agrupadas para formar categorías de electrofacies mediante el uso de una red de auto-organización neural.

Existen numerosos trabajos sobre la aplicación de las redes de neuronales aplicadas a la clasificación de facies, desde metodologías convencionales hasta metodologías híbridas (combinan sistemas neuronales con alguna otra técnica numérica) por citar algunas técnicas: están el uso de las redes de base radial o probabilísticas (González y Méndez, 2007), los sistemas híbridos de redes neuronales y lógica difusa (Ferraz y García, 2005), Redes neuronales y clasificadores de Markov (Mirowski y McCormick, 2003) y

conjuntos de técnicas como redes neuronales, análisis de agrupamiento y componentes principales (Ma, 2011). Sin embargo los modelos más utilizados para resolver el problema de clasificación de electrofacies, son las redes Neuronales de retropropagación (de estas el modelo perceptrón multicapa - MLP es el más utilizado) y los mapas autorganizados o redes de Kohonen.

En Alpana (2000), se utilizaron grupos de redes neuronales, trabajando en conjunto para la solución al problema de la estimación de litofacies, utilizando un modelo o esquema neuronal, llamado maquinas comité (Committee Machine), el cual aplica los enfoques de las redes neuronales de tipo modular (Modular Neural Network) y las redes neuronales de conjunto (Ensemble Neural Network), de las cuales se hablará con mayor detalle en este capítulo.

Existen diferentes paqueterías comerciales con numerosas técnicas para la estimación de electrofacies (Análisis de agrupamiento - Power Log, lógica difusa - Interactive Petrophysics, redes neuronales - Petrel - Statimagic - Geoview, etc), sin embargo es necesario plantear una metodología o un esquema estratégico para su aplicación. En el presente trabajo se pretende establecer una metodología para la estimación de electrofacies, utilizando tres modelos neuronales: mapas auto-organizados, mapas perceptrón multicapa y sistemas de combinación de redes neuronales.

2.2 Las Redes Neuronales Artificiales (RNA)

Las redes neuronales son tecnología que deriva de ciencias tales como la neurociencia, las matemáticas y la estadística, así como de la física, las ciencias computacionales y la ingeniería. Las redes neuronales artificiales han sido inspiradas en el comportamiento del cerebro humano, el cual es un sistema complejo de procesamiento de información, no lineal y paralelo (Haykin, 1999), que aprenden de la experiencia (Anderson y McNeill, 1992).

Una red neuronal como se mencionó anteriormente puede ser descrita como un procesador distribuido en paralelo y masivamente. Compuesta de simples unidades de procesamiento llamadas neuronas, la cual tiene una tendencia natural para almacenar conocimiento experimental y hacerlo disponible para su uso. Se asemeja al cerebro humano en los siguientes aspectos (Alpana, 2002):

1. El conocimiento es adquirido por la red a través del proceso de aprendizaje.
2. Las fortalezas de las conexiones inter-neuronales, conocidos como pesos sinápticos, son usado para almacenar los conocimientos adquiridos.

El cerebro humano tiene la capacidad de realizar ciertos cálculos, por ejemplo reconocimiento de patrones, la percepción, etc., de forma más rápido que una computadora. Ha habido muchos intentos de emular las funciones cerebrales con modelos de computadora y aunque ha habido algunos espectaculares, ha sido imposible a la fecha emular el complejo funcionamiento del cerebro humano (Alpana, 2002).

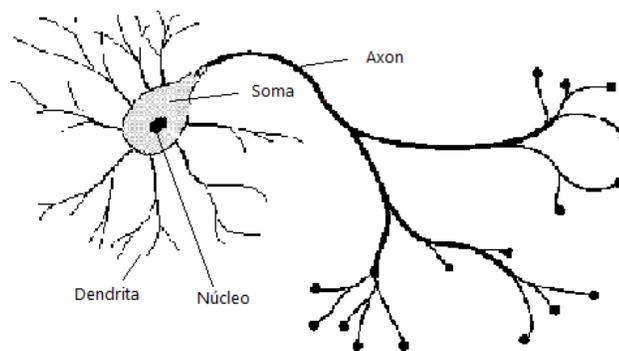


Fig. 2.1. Neurona Biológica. Editada de Alpana 2000.

La unidad celular fundamental del sistema nervioso del cerebro es una neurona, la cual es un simple elemento de procesamiento que combina señales de otras neuronas a través rutas de entrada llamadas dendritas. Si la señal de entrada combinada es lo suficientemente fuerte entonces la neurona reacciona, produciendo una señal de salida a lo largo del axón, el cual conecta a las dendritas de otras neuronas. La Fig. 2.1 es un esquema de una neurona biológica mostrando sus diversos componentes.

Cada señal llega a la neurona a lo largo de una dendrita y pasa a través de una sinapsis o una unión sináptica. Esta unión es un espacio infinitesimal en donde la dendrita es llenada con un fluido neurotransmisor que acelera o retarda el flujo de cargas eléctricas. Las acciones fundamentales de la neurona son químicas en la naturaleza y este fluido produce señales eléctricas que van a los núcleos o somas de las neuronas, los cuales son los encargados de procesar la información. El ajuste de la impedancia o conductancia del espacio sináptico es un proceso muy importante, ya que las fuerzas sinápticas de las neuronas están ajustadas para que el cerebro pueda aprender y almacenar información.

2.2.1 Modelo Neuronal

Similarmente a la neurona biológica, una neurona artificial es una unidad fundamental de una red neuronal artificial. El diagrama de bloque de la Fig. 2.2 muestra el modelo de una neurona. Sus básicos elementos son (Alpana, 2000):

- 1) Un conjunto de sinapsis o conexiones, cada una de las cuales está caracterizada por un peso propio. Una señal x_j en la entrada de la sinapsis j conectada a la neurona k , es multiplicada por el peso sináptico w_{kj}
- 2) Un elemento que suma las señales de entrada, ponderadas por la respectiva sinapsis de la neurona.
- 3) Una función de activación para limitar la amplitud de las salidas de una neurona. Limita el rango de amplitud permisible de una señal de salida a algunos valores finitos.
- 4) El modelo neuronal también incluye un factor aplicado externamente conocido como bias (o factor de polarización) b_k , el cual tiene el efecto de incrementar o disminuir la entrada de la red a la función de activación, dependiendo de si es positivo o negativo

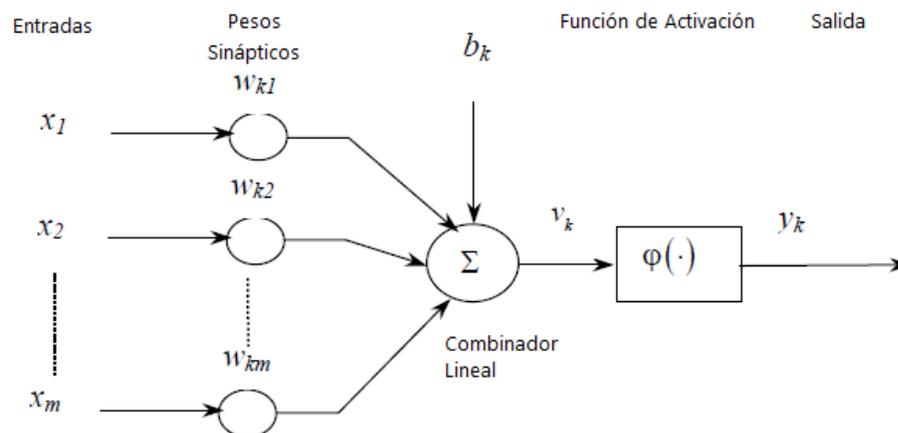


Fig. 2.2. Modelo de Neurona Artificial, Editada de Haykin, 1992

Matemáticamente puede ser descrita como (Haykin, 1999):

$$y_k = \varphi(u_k + b_k) \tag{2.1}$$

Donde y_k es la salida de la red, b_k es el bias y u_k es la combinación lineal de los pesos sinápticos w_{kj} y los valores de la entrada m dimensional x_j , definiéndose de la siguiente forma:

$$u_k = \sum_{j=1}^m w_{kj}x_j \tag{2.2}$$

La relación entre la combinación lineal u_k y el potencial de activación v_k es:

$$v_k = u_k + b_k \tag{2.3}$$

La función de activación $\varphi(v)$ define la salida de una neurona en términos del campo local inducido v_k . Las tres más comunes funciones de activación son:

i) Función Hardlimit

$$\varphi(v) = \begin{cases} 1 & \text{si } v \geq 0 \\ 0 & \text{si } v < 0 \end{cases} \tag{2.4}$$

La función Harlimit simétrica es descrita como:

$$\varphi(v) = \begin{cases} 1 & \text{si } v \geq 0 \\ -1 & \text{si } v \leq 0 \end{cases} \tag{2.5}$$

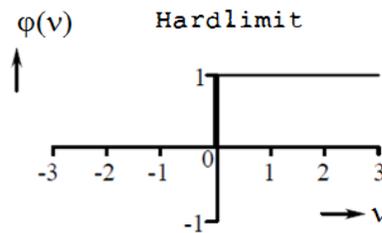


Fig. 2.3. Función Hardlimit simétrica

La función Harlimit simétrica es la más comúnmente se utiliza en reconocimiento de patrones.

ii) Función lineal.

En la función lineal $\varphi()$ es igual a salida de una función. Esta función de activación se usa en el reconocimiento de patrones y en problemas de aproximación.

$$\varphi(v) = v \tag{2.6}$$

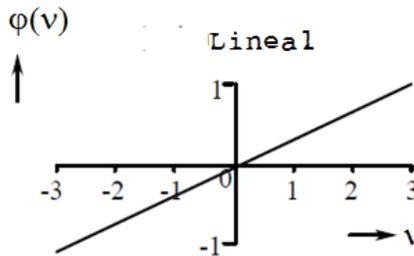


Fig. 2.4. Función Lineal.

iii) Función sigmoidea.

Esta es la forma de función de activación más común en la construcción de redes multicapa que son entrenadas con el algoritmo de retropropagación por que tiene la característica de ser no decreciente y diferenciable en el rango $0 \leq \varphi(v) \leq 1$

Un ejemplo de la función sigmoide es:

$$\varphi(v) = \frac{1}{[1 + e^{-\alpha v}]} \tag{2.7}$$

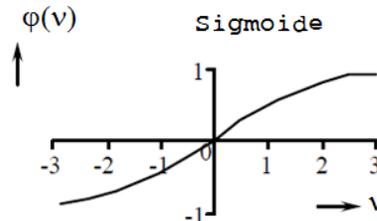


Fig. 2.5. Función Sigmoidea.

Donde α es el parámetro pendiente (slope parameter) de la función sigmoidea. La pendiente en el origen es igual a $\frac{\alpha}{4}$. En el límite cuando el parámetro pendiente se aproxima al infinito, la función sigmoidea se vuelve una función umbral. Una función sigmoidea tiene un rango de valores continuos de 0 a 1. Algunas veces es deseable tener la función de activación en un rango de -1 a 1 en el cual la función de activación asume una forma antisimétrica con respecto al origen. Entonces $\varphi(v)$ puede ser dada por una tangente hiperbólica definida como

$$\varphi(v) = \tanh(v) \tag{2.8}$$

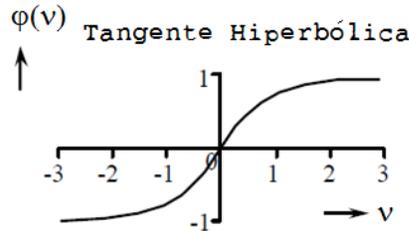


Fig. 2.6. Tangente Hiperbólica.

iv) Función Rampa

Se describe como:

$$\varphi(v) = \begin{cases} 1 & \text{si } v \geq \frac{1}{2} \\ v & \text{si } \frac{1}{2} > v > -\frac{1}{2} \\ -1 & \text{si } v \leq -\frac{1}{2} \end{cases} \tag{2.9}$$

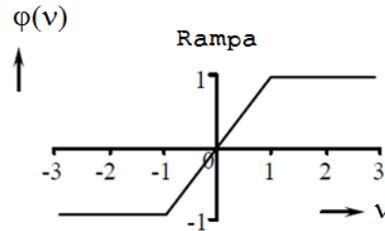


Fig. 2.7. Rampa

2.2.2 Procesamiento y extracción de características.

Debido a que la dimensión de las variables juega un papel fundamental en el rendimiento de la red, es habitual realizar técnicas para ayudar a aliviar problemas de dimensión, la cual consiste en combinar las variables de entrada para hacer un menor número de variables nuevas llamadas características. La extracción de características, es un tipo de pre-procesamiento, en el cual se transforman las variables de entrada a nuevas variables que poseen la información fundamental o característica que define al conjunto de datos de entrenamiento (Bishop, 1995).

A menudo es recomendable realizar un pre-procesamiento a los datos de entrada, con el fin de tener datos normalizados (en el intervalo de -1 a 1 o de 0 a 1), con respecto a los datos mínimo y máximo o en su defecto normalizar la media y la desviación estándar (Manual de Matlab 2009a y 2011b). En diversas aplicaciones, las salidas también experimentan un post-procesamiento para obtener alguna forma requerida a la salida de la red (principalmente en problemas de aproximación de funciones).

El conocimiento a priori sirve para definir las propiedades que tendrán las variables transformadas, por ejemplo la invariancia en la rotación o traslación en el tiempo, etc. La información a priori es aquella que se tiene acerca de la forma deseada de la solución, la cual es adicional a la información dada por los datos de entrenamiento. La inclusión de información a priori al diseño de un sistema de reconocimiento de patrones puede mejorar drásticamente su realización y el uso del pre-procesamiento es una vía importante para conseguirlo (Bishop, 1995).

2.3 Clasificación y tipos de redes Neuronales

Existen diversos criterios para catalogar los diferentes modelos neuronales. Pueden clasificarse de acuerdo con cuatro criterios básicos (Hilera y Martínez, 1995): (1) la naturaleza de las señales de entrada y salida, (2) la topología de la red, (3) el mecanismo de aprendizaje que utilizan y (4) el tipo de asociación de las señales de entrada-salida y la forma de representar estas señales.

2.3.1 De acuerdo con la naturaleza de los datos.

De acuerdo con la naturaleza de las señales de entrada y de salida, es decir el tipo de dato que son capaces de procesar, las redes neuronales se pueden clasificar en analógicas, discretas (generalmente binarias) e híbridas (Hilera y Martínez, 1995):

- a) Las redes continuas (analógicas) procesan datos de entrada de naturaleza analógica, los cuales son valores reales continuos, que comúnmente estarán normalizados y acotados en el intervalo $[-1,1]$ o en el $[0,1]$, para dar respuestas también continuas. Las redes analógicas suelen presentar funciones de activación continuas, habitualmente lineales o sigmoides. Entre esta categoría destacan las redes de entrenamiento Backpropagation, la red continua de Hopfield, la de Contra propagación, la de Memoria Lineal Asociativa, la red tipo Brain-State-in-Box, y las redes de aprendizaje competitivo, como los modelos de Kohonen (mapas auto-organizados o S.O.M.) y la red tipo L.V.Q. (Learning Vector Quantizer)

b) Las redes discretas (binarias) procesan datos de naturaleza discreta, habitualmente valores lógicos booleanos, en un intervalo $\{0,1\}$, para acabar emitiendo una respuesta discreta, en este caso, las funciones de activación de las neuronas serán del tipo escalón. Entre las redes binarias destacan la Máquina de Boltzman, la Máquina de Cauchy, la red discreta de Hopfield, el Cognitrón y el Neogognitrón.

c) Las redes híbridas procesan entradas analógicas para dar respuestas binarias. Dentro de este tipo de modelos destacan el Perceptrón, la red Adaline y la Madaline.

2.3.2 De Acuerdo con su topología:

La topología o arquitectura de una red neuronal es la forma en que se organizan y disponen las neuronas en la red formando capas o agrupaciones de neuronas (Hilera y Martínez, 1995).

Desde el punto de vista de la arquitectura, los parámetros fundamentales de una red neuronal son los siguientes:

- Número de capas.
- Número de neuronas por capa.
- Grado de conectividad.
- Tipo de conexiones entre neuronas.

Se denomina capa o nivel a un conjunto de neuronas cuyas entradas provienen de la misma fuente (que puede ser otra capa de neuronas) y cuyas salidas se dirigen al mismo destino (que puede ser otra capa de neuronas) (Hilera y Martínez, 1995) y dependiendo de su situación dentro de la red, se pueden distinguir tres tipos de capas:

- Capa de entrada. Es la capa que recibe directamente la información o datos de entrada provenientes de las fuentes externas a la red.
- Capas ocultas. Están formadas por neuronas ocultas. Son capas internas a la red y no tienen contacto directo con el entorno exterior. El número de capas ocultas presentes en una red puede variar desde cero (Mapas auto-organizados) hasta un número elevado.
- Capa de salida. La función de esta capa es transferir información de la red hacia el exterior.

En una clasificación en términos topológicos las primeras distinciones a establecer son respecto al número de capas de la red, diferenciando entre las redes monocapa y multicapa (2, 3, etc.) y el patrón de conexiones que presenta, distinguiendo entre unidireccionales o recurrentes.

Las redes unidireccionales son aquellas que no presentan ciclos, es decir, no se puede trazar un camino de una neurona a sí misma (Artiaga y Escobar, 2006). Si el sentido de la dirección es progresivo (Fig. 2.7a), es decir, existe una dirección en sentido jerárquico de la entrada a la salida de la red, se les denomina redes de *alimentación hacia delante* (feedforward), donde la dirección del procesamiento es única. Si las redes presentan al menos un ciclo, en donde la salida de una neurona se conecta con su propia entrada o a la entrada de neuronas de capas anteriores, son llamadas redes recurrentes (Bishop, 1995) (Fig. 2.9).

2.3.2.1 Las redes Monocapa.

Las redes monocapa son aquellas que sólo cuentan con una capa de neuronas oculta. Dentro de las redes monocapa unidireccionales con alimentación hacia adelante, se encuentran las redes perceptrón,

ADALINE y MADALINE. Dentro de las redes de conexiones laterales (Fig. 2.8b), se encuentran los modelos de la red de Hopfield, la red Brain-State-in-Box, la máquina de Boltzman, la máquina de Cauchy y los mapas autoorganizados o red Kohonen. (Gutiérrez, 2000).

En el rubro de las redes monocapa recurrentes, tenemos por ejemplo los modelos Brain-State-in Box recurrente, la red de Elman, la red de Hopfield y la Máquina de BoltzmRNA recurrentes. Las redes monocapa recurrentes se utilizan típicamente en tareas relacionadas con lo que se conoce como autoasociación.

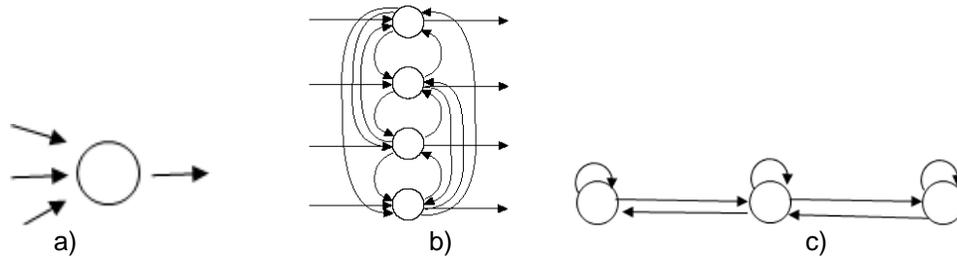


Fig. 2.8 Tipo de conexiones en una red monocapa (Modificado de Fernández et. al., 2008; e Hilera y Martínez 1995).
 a) Red monocapa con conexión de alimentación hacia adelante (feedforward).
 b) Red Monocapa con conexiones laterales.
 c) Red Monocapa con conexiones laterales y autorecurrentes.

2.3.2.2 Las redes Multicapa.

Las redes multicapa son aquellas que disponen de un conjunto de neuronas jerarquizadas en distintos niveles o capas y son especialmente útiles en aplicaciones de reconocimiento o clasificación de patrones (Castillo, 2007).

Las redes multicapa tienen una capa de entrada, otra de salida, y eventualmente dos o varias capas intermedias u ocultas (Fig. 2.9). Normalmente, todas las neuronas de una capa reciben señales de entrada de otra capa anterior y envían las señales de salida a una capa posterior (en un sentido de entrada hacia salida), es decir, presentan conexiones de alimentación hacia adelante (feedforward) (Fig. 2.9 y 2.10a), aunque pueden existir conexiones a neuronas de cualquiera de las capas anteriores (feedback o conexiones recurrentes de alimentación hacia atrás). Dentro de estos modelos los más conocidos son: la red de memoria adaptativa lineal o LAM (Linear Adaptative Memory), la red Drive-Reinforce-Ment y la red perceptrón multicapa (Multilayer perceptrón o también llamada red Backpropagation, por el algoritmo de entrenamiento utilizado en este tipo de arquitectura).

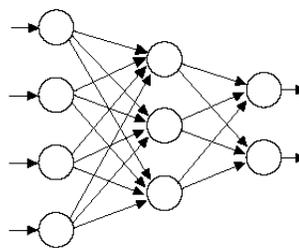


Fig. 2.9 Red Neuronal con tres capas ocultas y conexiones de alimentación hacia adelante. (Modificado se Fernández et. al., 2008)

Dentro de los principales modelos con conexiones alimentación hacia adelante y recurrentes, se encuentran el Cognitron, Neocognitron, la red ART (Adaptive Resonance Theory), la red BAM (Bidirectional Associative Memory) y las máquinas multicapa de Boltzman y Cauchy (Villanueva, 2002; Flores y Fernández et.al, 2008).

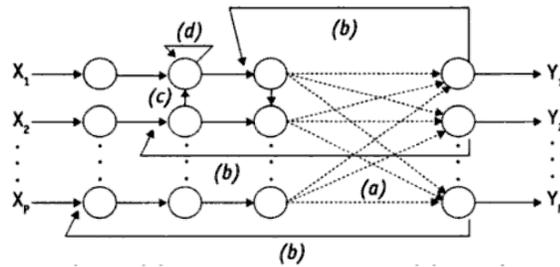


Fig. 2.10 Red multicapa con diferentes tipos de conexiones (Tomado de Fernández et. al)

- a) Conexiones de alimentación hacia adelante (feedforward).
- b) Conexiones recurrentes de alimentación hacia atrás (feedback).
- c) Conexiones laterales.
- d) Conexiones autorecurrentes.

2.3.3 Tipo de Aprendizaje.

Podemos definir el aprendizaje de una RNA como el proceso por el cual una red neuronal modifica sus pesos en respuesta a una información de entrada (Hilera y Martínez, 1995). Los cambios que se producen durante el proceso de aprendizaje se reducen a la eliminación, modificación y creación de conexiones entre las neuronas. La creación de una nueva conexión implica que el peso asociado a la misma pasa a tener un valor distinto de cero; de la misma forma, una conexión se elimina, cuando su peso pasa a valer cero. Se puede afirmar que este proceso ha terminado (la red ha “aprendido”), cuando los valores de los pesos permanecen estables.

Un aspecto importante respecto al aprendizaje de las redes neuronales es el conocer cómo modifican los valores de los pesos; es decir, cuáles son los criterios que se siguen para cambiar el valor asignado a las conexiones cuando se pretende que la red “aprenda” una nueva información (Villanueva, 2002). Estos criterios determinan lo que se conoce como la *regla de aprendizaje* de la red. Por consiguiente, se puede establecer una clasificación de las redes neuronales desde el punto de vista del tipo de aprendizaje utilizado; de esta manera, atendiendo a esta clasificación, podemos distinguir entre tres tipos de redes:

- a. Redes neuronales con aprendizaje supervisado
- b. Redes neuronales con aprendizaje no supervisado.
- c. Redes neuronales con aprendizaje híbrido

La diferencia fundamental entre estos tipos de redes radica en la existencia o no de un agente externo (supervisor) que controle el proceso de aprendizaje de la red.

Otro criterio que se puede utilizar para diferenciar las reglas de aprendizaje es (Zapata y Díaz, 2007):

- Redes neuronales con aprendizaje *on-line*.
- Redes neuronales con aprendizaje *off-line*.

El aprendizaje *off-line* implica que existen dos tipos de fases asociadas a la red: una *fase de aprendizaje* o *entrenamiento* y una *fase de operación* o *funcionamiento*. Asimismo, debe existir un conjunto de datos de entrenamiento y un conjunto de datos de test o prueba que serán utilizados en la correspondiente fase para comprobar y evaluar el funcionamiento de la red (Gutiérrez, 2000).

En las redes con aprendizaje *on-line* no se distingue entre fase de entrenamiento y fase de operación, de tal forma que los pesos varían dinámicamente siempre que se presente una nueva información al sistema (Gutiérrez, 2000).

2.3.3.1. Redes con aprendizaje supervisado

El aprendizaje supervisado se caracteriza porque el proceso se realiza mediante un entrenamiento controlado por un agente externo (guía, supervisor, maestro) que determina la respuesta que debería generar la red a partir de una entrada determinada. El supervisor comprueba la salida de la red y en el caso de que ésta no coincida con la deseada, se procederá a modificar los pesos de las conexiones, con el fin de conseguir que la salida obtenida se aproxime a la deseada. En este tipo de aprendizaje se suele considerar, tres formas de llevarlo a cabo, que dan lugar a los siguientes tipos de aprendizajes (Gutiérrez, 2000).

- a) **Aprendizaje por corrección de error.** Consiste en ajustar los pesos de las conexiones de la red en función de la diferencia entre los valores deseados y los obtenidos en la salida de la red, es decir, en función del error cometido en la salida. Un ejemplo de este tipo de algoritmo lo constituye la regla de aprendizaje del Perceptrón.
- b) **Aprendizaje por refuerzo.** Es un aprendizaje con características del supervisado y con características del auto-organizado. No se proporciona una salida deseada, pero sí se indica a la red el error que comete (Correa y Porras, 1999). Un ejemplo de algoritmo por refuerzo lo constituye el denominado *Linear Reward-Penalty* ó *LRP* (algoritmo lineal con recompensa y penalización), presentado por Narendra y Thathacher en 1974.
- c) **Aprendizaje estocástico.** Este tipo de aprendizaje consiste básicamente en realizar cambios aleatorios en los valores de los pesos de las conexiones de la red y evaluar su efecto a partir de distribuciones de probabilidad y del objetivo deseado (Gutiérrez, 2000). Una red que utiliza este tipo de aprendizaje es la conocida como Máquina de BoltzmRNA, ideada por Hinton, Ackley y Sejnowski en 1984, que combina el aprendizaje Hebbiano o con aprendizaje por corrección de error.

2.3.3.2. Redes con aprendizaje no supervisado

Las redes con aprendizaje no supervisado (también conocido como auto-supervisado o auto-organizado) no requieren influencia externa para ajustar los pesos de las conexiones entre sus neuronas (Gutiérrez, 2000). La red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta. Por ello, suele decirse que estas redes son capaces de auto-organizarse.

Existen varias posibilidades en cuanto a la interpretación de la salida de estas redes, que dependen de su estructura y del algoritmo de aprendizaje empleado. En algunos casos, la salida representa el grado de *familiaridad* o similitud entre la información que se le está presentando en la entrada, realizando así un establecimiento de categorías o *clustering*. Este términos prácticos este tipo de entrenamiento es un *mapeo de las características más representativas de los datos*, realizando una *codificación* de los datos de entrada, generando a la salida una versión codificada de la entrada, pero manteniendo la información relevante de los datos (muy similar al método de las componentes principales).

En cuanto a los algoritmos de aprendizaje no supervisado, en general se suelen considerar los dos siguientes:

- a) *Aprendizaje Hebbiano.*

Este tipo de aprendizaje se basa en un postulado formulado por Donald O. Hebb en 1949. El aprendizaje Hebbiano consiste básicamente en el ajuste de los pesos de las conexiones de acuerdo con la correlación de los valores de activación (*U. de G, 2012 y Gutiérrez, 2000*). Este tipo de aprendizaje fue empleado por Hopfield en la red que lleva su nombre (Red de Hopfield), introducida

en 1982. Grossberg también utilizó este tipo de aprendizaje en las redes llamadas Additive Grossberg y Shunting Grossberg, (Villanueva, 2012).

b) *Aprendizaje competitivo y cooperativo.*

En las redes con aprendizaje competitivo y/o cooperativo, se suele decir que las neuronas compiten y/o cooperan unas con otras con el fin de llevar a cabo una tarea dada. Con este tipo de aprendizaje, se pretende que cuando se presente a la red cierta información de entrada, sólo una de las neuronas de salida de la red, o una por cierto grupo de neuronas, se active. Por tanto, las neuronas compiten por activarse, quedando finalmente una, o una por grupo, como neurona vencedora, quedando anulado el resto (Diazaraque, 2012). El objetivo de este aprendizaje es categorizar los datos (formar grupos o clústeres). De esta forma, la información que pertenece a la misma categoría, debe activar la misma neurona de salida. Un ejemplo de este tipo de aprendizaje es el desarrollado por Kohonen, conocido como *Learning Vector Quantization* (LVQ).

El aspecto geométrico de la disposición de las neuronas de una red también es la base de un caso particular de aprendizaje competitivo introducido por Kohonen en 1982, conocido como *feature mapping* (mapeo de características), aplicado en redes con una disposición bidimensional de las neuronas de salida, que permiten obtener mapas topológicos en los que, de algún modo, estarían representadas las características principales de la información presentada a la red.

2.3.4 Tipo de asociación de las señales de entrada y salida

Las redes neuronales artificiales almacenan cierta información “aprendida”, la cual se registra de forma distribuida en los pesos asociados a las conexiones entre neuronas (Gutiérrez, 2000; Ochoa, 2003).

Existen dos formas primarias de realizar esta asociación entre la entrada y la salida, que depende la naturaleza de la información almacenada en la red. Una primera sería la denominada *hetero-asociación*, que se refiere al caso en el que la red “aprende” parejas de datos $[(A_1, B_1), (A_2, B_2), \dots, (A_N, B_N)]$, de tal forma que cuando se presenta cierta información de entrada A_i , deberá responder generando la correspondiente salida asociada B_i . La segunda se conoce como *auto-asociación*, donde la red “aprende” ciertas informaciones A_1, A_2, \dots, A_N , de tal forma que cuando se le presenta una información de entrada, se realizará una autocorrelación, respondiendo con uno de los datos almacenados, el más parecido al de la entrada (Zapata y Díaz, 2007).

Los dos mecanismos de asociación expuestos arriba dan lugar a una nueva clasificación de las redes neuronales (ver Tabla 2.1)

Una *red heteroasociativa* podría considerarse aquella que calcula cierta función, que en la mayoría de los casos no podrá expresarse analíticamente, entre un conjunto de entradas y un conjunto de salidas, correspondiendo a cada posible entrada una determinada salida. Por otra parte, la *red autoasociativa* tiene como principal misión reconstruir una determinada información de entrada que se presenta incompleta o distorsionada (asocia el dato almacenado más parecido) (Hilera y Martínez, 1995).

Tabla 2.1. Clasificación de algunos modelos neuronales, en función del tipo de asociación de los datos de entrada y de salida. Modificado Fernández et. al

Redes heteroasociativas		Redes Autoasociativas
Perceptron	Linear Associative Memory	Brain-Satate-in-a-box
Adeline/Madaline	Optimal Linear Associative Memory	Hopfield
Backpropagation	Drive-Reinforcement	Optimal Linear Associative Memory
Linear Reward Penalty	Fuzzy Associative Memory	Additive Grossberg
Associative Reward Penalty	Counterpropagation	Shuting Grossberg
Adaptive Heuristic Critic	Bidirectional Associate Memory (BAM)	Self Organising Feature Map (SOFM)
Boltzmann Machine	Adaptive BAM	
Cauchy ;achine	Cognitrón/Neocognitrón	
Learning Matrix	Learning Vector Quantizer	
Temporal Associative Memory	Adaptative Resonance Theory	

2.3.5 Por su aplicación

Las redes neuronales pueden clasificarse en función de los problemas que resuelven más comúnmente (Hilera y Martínez, 1995); por citar algunos se tiene:

- a) *Clasificación o reconocimiento de patrones.*- Proceso en el que se decide a qué clase pertenece (generalmente definida previamente) cada punto de una serie de datos dada.
- b) *Asociación.*- Proceso de recuperación de un dato a partir de una entrada relacionada con el dato almacenado. La entrada puede ser una versión incompleta o deformada del dato buscado.
- c) *Agrupamiento (Clustering).*- Se presenta cuando no se conoce una clasificación de los datos de entrada y se espera que la red genere esta clasificación (generación de prototipos).
- d) *Optimización.*- Cuando se desea obtener la mejor respuesta de un sistema, cuyos parámetros han sido establecidos previamente. Las RNA proporcionan un procedimiento rápido para generar soluciones a los problemas de optimización.
- e) *Predicción y Control.*- Cuando se desea estimar una función o realizar una predicción cuando se desea anticipar el estado futuro de un sistema.

2.3.5.1 Reconocimiento de Patrones.

El reconocimiento de patrones se define formalmente como el proceso mediante el cual un patrón o señal recibida se asigna a una clase de un número prescrito de clases (Alpana, 2002). Una red neuronal realiza el reconocimiento de patrones, pasando por una sesión de entrenamiento, durante la cual la red es presentada en repetidas ocasiones a un conjunto de patrones de entrada junto con la categoría a la que cada patrón en particular pertenece (epochs o número de iteraciones)¹. Más tarde un nuevo patrón se presenta a la red que no se ha visto antes, pero que pertenece a la misma población de los patrones utilizados para entrenar la red. La red es capaz de identificar la clase de un patrón en particular, debido a la información extraída de los datos de entrenamiento. El reconocimiento de patrones realizado por una red neuronal es de naturaleza estadística, con patrones representados por puntos en un espacio

¹ Iteración: Proceso en el cual un valor del conjunto de datos es presentado a la red.

Epoch: Conjunto de iteraciones que finaliza cuando todos los valores del conjunto de datos de entrenamiento han sido presentados a la red, durante el proceso de aprendizaje

multidimensional, el cual, se divide en regiones y cada una de estas se asocia con una clase y las fronteras de decisión son determinadas por el proceso de entrenamiento.

Como se mencionó en la sección 2.2.2, la extracción de características es el proceso de transformar los patrones de entrada de un espacio dimensional, en variables de otro espacio de menor dimensión y así facilitar la tarea de clasificación. Entonces la clasificación puede ser entendida también como el proceso por el cual las características son mapeadas dentro de una clase, en un espacio n-dimensional, donde n es el número de clases a ser distinguidas. Esta tarea puede ser realizada usando una red supervisada o no supervisada.

2.3.5.2 Aproximación de funciones

Las redes neuronales tienen la capacidad de hacer una aproximación de la función entre las entradas y las salidas con el grado de precisión deseado. Considere un mapeo entrada-salida descrito por la relación funcional:

$$d = f(x) \quad 2.10$$

La función $f(x)$ es asumida desconocida. Si damos un conjunto de muestras para mapear la función.

$$D = \{(x_i, d_i)\}_{i=1}^m \quad 2.11$$

Se tiene que diseñar una red neuronal que aproxime la función desconocida $f(x)$, mediante la función $F(x)$ que describe el mapeo entrada y salida. $F(x)$ debe encontrarse lo suficientemente cercana a $f(x)$ para todas las entradas tales que:

$$F(x) - f(x) < \varepsilon \quad 2.12$$

Donde ε es un número positivo pequeño.

Dado que el tamaño m del conjunto de entrenamiento es lo suficientemente grande y la red está equipada con un adecuado número de parámetros libres (pesos y bias) entonces la aproximación del error ε puede ser lo suficientemente pequeña, por lo tanto $F(x)$ será una buena aproximación de $f(x)$

2.4 Arquitectura perceptrón.

Como se mencionó anteriormente, Rosenblatt estudió las redes de capas simples con funciones de activación en 1962, las cuales llamó perceptrones y fueron utilizadas en la solución de problemas de clasificación en las que las entradas y salidas eran binarias, como en el reconocimiento de caracteres (Bishop, 1995). La revisión de las propiedades del perceptrón, fueron revisadas en Block (1962).

La arquitectura perceptrón es la más simple de todas las redes neuronales, ya que consta de una neurona de salida (Fig. 2.2) y puede clasificar vectores en dos categorías. Esta es la limitación de la arquitectura perceptrón. Los pesos de la red son ajustados por un algoritmo de aprendizaje supervisado (Alpana, 2002). El perceptrón calcula una suma ponderada (Ec. 2.2), la cual entra a una función de activación (ver ecuaciones 2.4 y 2.5), la cual, dependiendo del valor umbral, generará una salida, cuyo valor es 0 o 1 (Anderson y McNeill, 1992).

$$y = \text{hardlim}(W^T x + b) \quad 2.13$$

El algoritmo de entrenamiento del perceptrón, es de tipo supervisado, donde los pesos de la red se actualizan para ajustar la salida de una entrada dada, a una salida esperada (supervisor). Sea D un conjunto de entrenamiento consistente de n entradas y salidas tales que si d es la salida deseada (supervisor) e y es la salida calculada, entonces el error para la iteración k_{th} está dada por:

$$e(k) = d(k) - y(k) \tag{2.14}$$

Los ajustes a los pesos y el bias están dados por:

$$w(k + 1) = w(k) + e(k)x(k) \tag{2.15}$$

y

$$b(k + 1) = b(k) + e(k) \tag{2.16}$$

La regla de aprendizaje perceptrón es simple, poderosa y converge siempre a los pesos que cumplen con la clasificación deseada (Alapana, 2002). Sin embargo, la limitación es que solo puede clasificar los vectores de entrada que puedan ser separados por una frontera lineal, es decir que sean linealmente separables. Como ejemplo de lo anterior, se tienen los problemas que representan las funciones lógicas OR, AND y XOR, las cuales, son consideradas, como un problema geométrico de clasificación binaria. La Tabla 2.2 muestra las salidas de las funciones AND, OR y XOR (donde 1 es cierto y 0 es falso) para dos variables lógicas M y N.

Tabla 2.2. Funciones OR, AND y XOR, para dos variables lógicas M y N.

Punto	Variable		Función		
	M	N	And	Or	Xor
A	0	0	0	0	0
B	1	0	0	1	1
C	1	1	1	1	0
D	0	1	0	1	1

En la Fig. 2.11, se muestran las fronteras de decisión de las funciones OR y AND, donde es posible dibujar la línea OR que separa los puntos B, C y D, del punto A, el cual, es el único valor falso (0) en la Tabla 1.2, para la función OR. De igual forma la línea AND, separa los puntos A, D y B, del punto C, el cual, toma el único valor cierto (1) en la función AND, mostrada en la Tabla 2.2.

La función XOR tendría que tener a los puntos A y C a un lado de la línea y a los puntos B y D del otro lado, por lo que es claro que no hay una línea recta que pueda separarlos de tal manera, que este problema se considera que no es linealmente separable (Dunne, 2007).

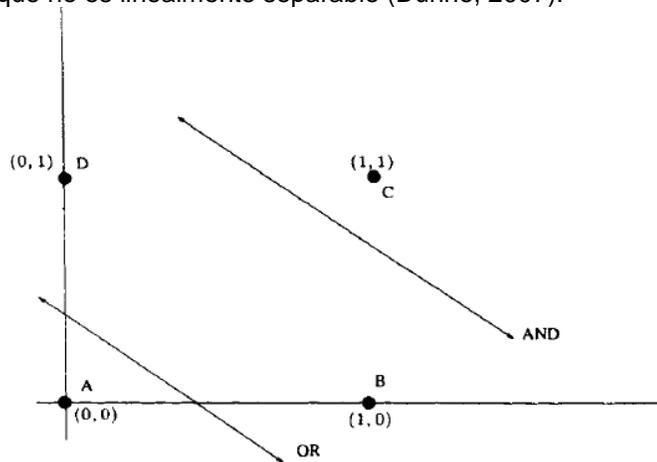


Fig. 2.11 Fronteras de decisión de las funciones OR y AND (tomado de Dunne, 2007).

2.5 Arquitectura ADALINE.

Bernard Widrow en 1960, introdujo la red ADALINE (ADaptive Llinear NEuron) la cual es muy similar al perceptrón. La función de transferencia es lineal a diferencia del perceptrón y la salida de la red está dada por.

$$y = \text{pureline}(W^T x + b) \quad 2.17$$

Donde la frontera de decisión es igual a la de la arquitectura perceptrón. El algoritmo de aprendizaje se conoce como la regla delta o algoritmo cuadrático medio (*LMS*), el cual también es un ejemplo de entrenamiento supervisado. El algoritmo ajustará los pesos y los bias de la arquitectura ADALINE minimizando el error cuadrático medio entre la salida objetivo y la salida calculada. Este error está definido por la siguiente función:

$$E[e^2] = E[(d - y)^2] = E[(d - W^T x)^2] \quad 2.18$$

El algoritmo *LMS* encuentra el punto mínimo del error cuadrático medio usando el algoritmo de descenso rápido o descenso por gradiente (Hagan et. Al 1996). Los ajustes de los pesos y los bias para la iteración K_{th} , usando el algoritmo *LMS*, es dado por:

$$w(k + 1) = w(k) + 2\eta e(k) x(k) \quad 2.19$$

$$b(k + a) = b(k) + 2\eta e(k) \quad 2.20$$

Donde η es una constante conocida como factor de aprendizaje.

Las redes ADALINE y el perceptrón tienen la misma limitación ellas solo podrían clasificar problemas linealmente separables, sin embargo el algoritmo *LMS* es más poderoso que la regla de aprendizaje del perceptrón. El perceptrón garantiza convergencia a la solución que correctamente categoriza los patrones de entrenamiento, la red resultante puede ser sensitiva a ruido ya que los parámetros yacen cerca de las fronteras de decisión (Alpana, 2002). El algoritmo *LMS* minimiza el error cuadrático medio, por lo tanto intenta mover la frontera de decisión lo más lejos de los patrones de entrenamiento como sea posible. El algoritmo tiene muchas más aplicaciones que la regla de aprendizaje del perceptrón, particularmente dentro del área de procesamiento de señales.

Se visualiza la relación que existe entre los parámetros libres (los pesos w y los bias b) y el error cuadrático medio que genera cada combinación de parámetros, se obtendrá una superficie de error en la cual la configuración de los pesos y los bias, producirá para cada punto un error (Mejía, 2004). El algoritmo *LMS* es óptimo para una sola neurona y problemas linealmente separables, porque la superficie del espacio del error cuadrático medio para una simple neurona tiene solo un punto mínimo y curvatura constante, lo cual da una solución única (Alpana, 2002).

Para una red perceptrón multicapa, por otro lado, la superficie de error puede tener muchos mínimos locales (Fig. 2.12) y la curvatura puede variar ampliamente en diferentes regiones del espacio de parámetros, de tal modo que el algoritmo *LMS* falla al producir una solución única en redes multicapas, ya que no se especifica como ajustar los pesos para unidades de las capas interiores durante el aprendizaje. En redes multicapa esto se conoce como problema de asignación de crédito (The credit assignment problem), según el cual, la cantidad de neuronas en la capa de entrada y capas intermedias, así como su distribución, dificulta la adecuada asignación de pesos debido a la limitación de los propios algoritmos de entrenamiento (Martínez, 2006-2007), ya que los algoritmos de aprendizaje en las capas internas no se extienden a capas posteriores, lo que dificulta la contribución del error de cada neurona al error final (McAuley y Simon, 1999). El problema de asignación de crédito fue resuelto usando el método de retropropagación del error (Hagan et. Al, 1996) el cual es una generalización del algoritmo *LMS*.

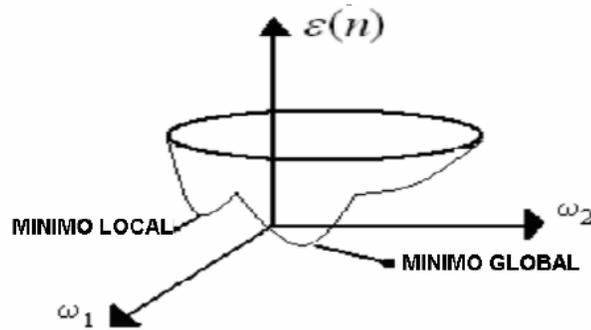


Fig. 2.12 Superficie de error esquemática para una red perceptrón multicapa (tomado de Mejía, 2004)

2.6 Red perceptrón multicapa.

Las redes perceptrón multicapas (MPL), son en la actualidad las redes más ampliamente usadas, pueden hacer la clasificación de patrones teniendo fronteras de decisión no linealmente separables, ya que la red consiste de muchas neuronas, distribuidas en las diferentes capas, Alpana (2002). Consiste de una capa de entrada, una o más capas internas (ocultas) y una capa de salida (Fig. 2.13). Este tipo de modelos comúnmente también son llamados redes multicapas de alimentación hacia adelante (feedforward) o redes backpropagation (por el algoritmo de entrenamiento). Las redes perceptrón multicapa se encuentran en la categoría de las redes de aprendizaje supervisado.

En las redes perceptrón multicapa, los vectores de entrada x valuados en el espacio real m dimensional, son presentados a cada una de las k neuronas de las capas ocultas a través del vector de pesos w . La neurona k que recibe una entrada j a través del peso sináptico w_{kj} (donde $k=1,2,3,\dots,h$ y $j=1,2,3,\dots,m$), calculará un valor de salida que es función de las entradas x_j y los pesos w_{kj} . Este valor de salida es presentado como la entrada de las neuronas de la capa siguiente y así de manera sucesiva. Al final de todo el proceso, se obtiene el valor de salida de la red, lo cual permite calcular el error entre dicho valor y el valor del objetivo (valor del maestro, guía o también llamado patrón de entrenamiento de salida). El algoritmo de aprendizaje que minimiza este error se denomina algoritmo de retropropagación (Backpropagation); el cual se publicó a mediados de los 80's para perceptrones multicapas (Alpana, 2002).

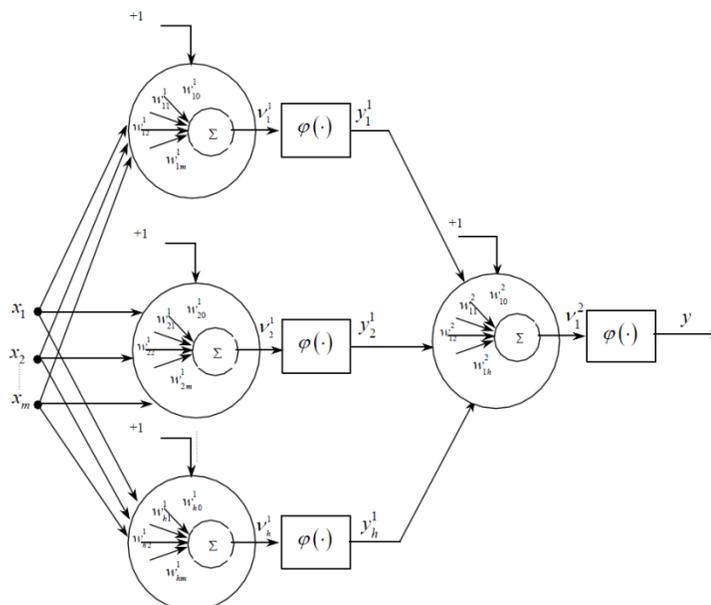


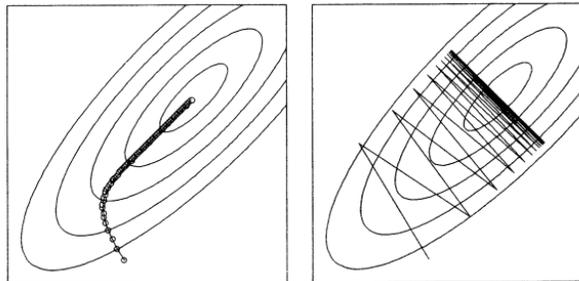
Fig. 2.13 Arquitectura de la red perceptrón multicapa (tomado de Alpana, 2002).

Hornik et. Al (1989) sugirió que si se tiene un número suficiente de capas ocultas, entonces la red MLP con una función sigmoide en la capa oculta y una función de transferencia lineal en la capa de salida pueden aproximar cualquier función a cualquier grado de precisión.

El algoritmo de retropropagación (*BP*) es la generalización del algoritmo *LMS* (least Mean square) y utiliza el error cuadrático medio como función de costo² o función de error. Dicha función está definida en el espacio multidimensional de pesos, para un conjunto dado de datos de entrenamiento. El espacio de pesos es una superficie que tendrá muchos mínimos (globales y locales) y la regla de aprendizaje buscará el punto donde se encuentra el mínimo global de esa superficie, a través de una trayectoria (generada durante el entrenamiento).

En el algoritmo *BP*, el error cuadrático medio se minimiza iterativamente usando el algoritmo de descenso rápido por gradiente³ (steepest descent algorithm). La diferencia entre el algoritmo *BP* y el *LMS* yace en la forma en la cual el gradiente del error es calculado. El algoritmo *BP* usa la regla de la cadena para obtener las derivadas del error cuadrático medio con respecto a los pesos y los bias en las capas ocultas. Las derivadas del error cuadrático medio son calculadas primero en la última capa de la red y entonces propagadas hacia las capas ocultas, usando la regla de la cadena, para así actualizar los pesos de la red. Al final de la actualización de los pesos se procede a calcular una nueva salida y un nuevo valor de la función de costo. Este proceso se repite un cierto número de veces, para cada elemento del conjunto de entrenamiento (datos del vector de entrada utilizados durante el entrenamiento ó también llamado patrones de entrenamiento de entrada) hasta que el error cuadrático medio total de la salida converja a un mínimo o hasta que algún límite es alcanzado, en el número de iteraciones establecidas durante el entrenamiento (Alpana. 2002). El algoritmo *BP* se utiliza solo para la fase de aprendizaje, mientras que la alimentación hacia adelante (feedforward) se utilizan tanto para el aprendizaje como para la operación de la red.

Uno de los mayores problemas con el algoritmo *BP* ha sido el largo tiempo de entrenamiento debido a que el método de descenso por gradiente es sencillo pero lento (Alpana, 2002). La tasa de aprendizaje (*TA*) es el parámetro que establece que tan rápido se efectúa el proceso de aprendizaje y que tan efectivo es (García, 2011), dicho parámetro es sensitivo a los cambios en los pesos. Cuanto menor sea la tasa de aprendizaje, menores serán los cambios en los pesos sinápticos de una iteración a la siguiente y más suavizada será la trayectoria de convergencia del espacio de los pesos. Por otro lado, si la tasa de aprendizaje se elige demasiado grande con el fin de acelerar el proceso de entrenamiento, los cambios en los pesos sinápticos serán mayores y harán inestable a la red (Fig. 2.14).



a) *TA* pequeña - Trayectoria Suave b) *TA* pequeña Trayectoria inestable

Fig. 2.14 Trayectoria en el espacio de los pesos sinápticos

² Dentro de la teoría de la estimación bayesiana, las funciones de costo dependen del error asociado a una estimación y son utilizadas para minimizar el riesgo bayesiano o costo medio (Teoría de la Señal y Comunicaciones, 2012)

³ El algoritmo de descenso rápido es un método utilizado en el cálculo de gradientes y su principal aplicación es la de resolver integrales complejas (Wang, 2008)

Con el fin de acelerar la convergencia del algoritmo de *BP*, junto con una estabilidad mejorada, se añade un término de momento a la actualización de pesos del algoritmo *BP*. El término momento es simple de implementar y aumenta significativamente la velocidad de convergencia. La inclusión del término momento, representa una pequeña modificación a la actualización de pesos y también puede tener la ventaja de impedir que la solución termine en mínimos locales de la superficie del espacio de error, durante el proceso de aprendizaje (Alpana, 2002).

El segundo método para acelerar la convergencia del algoritmo *BP* es utilizando el algoritmo Levenberg-Marquardt *BP* (*LMBP*) (Hagan et al., 1996). Está basado en el método de optimización de Newton y difiere del algoritmo usual *BP* en la manera en que las derivadas resultantes se usan para actualizar los pesos. El principal inconveniente del algoritmo es la necesidad de una gran cantidad de memoria y espacio de almacenamiento de los parámetros libres en la computadora. Si la red tiene más que unos miles de parámetros libres, el algoritmo se vuelve impráctico en máquinas comunes (Alpana, 2002).

Mientras se entrena una red MLP por el algoritmo *BP*, se calculan los pesos sinápticos esperando que la red pueda generalizar. La generalización es un concepto que implica que la red no memoriza los datos de entrada (cuando una red neuronal memoriza solo reconoce los datos con que fue entrenada), sino que le permite reconocer patrones de datos ligeramente diferentes y que nunca han sido presentados a la red. Se dice que una red generaliza bien cuando el mapeo entre la entrada y la salida es correcto para los patrones de prueba, los cuales son desconocidos para la red.

Se dice que una red particulariza (no generaliza bien), cuando termina memorizando los datos de entrada y lo anterior generalmente sucede, cuando la red es incapaz de filtrar características presente en los datos de entrenamiento y que no son propios la de función intrínseca⁴ de los datos que se van a modelar (como el ruido); a este fenómeno se le conoce como sobreajuste (Overfitting) o sobre entrenamiento (Overtraining), el cual, es el resultado de un mayor número de neuronas ocultas que las que son realmente necesarias, lo que contribuye a que la información no deseada en el espacio de entrada, por ejemplo el ruido, se almacene en los pesos sinápticos, respondiendo principalmente a las características del ruido, lo cual implica que la generalización está en función del número de parámetros libres.

Un número excesivo de parámetros libres, se traduce en una mayor dimensión en los espacios de las variables entrada, de pesos y de los bias, produciendo el efecto conocido como maldición de la dimensionalidad (Bishop, 1995), el cual ilustra que mientras más grande sea la dimensión de las variables implicadas en un sistema de clasificación, menor será el desempeño del clasificador. En la Fig. 2.15, se muestra el ejemplo de una función que ha sido muestreada y ajustada por polinomios de dimensiones diferentes. En la Fig. 2.15b el grado del polinomio es mayor que el de la Fig. 2.15a y su ajuste es mayor en los puntos muestreados, sin embargo el polinomio de la Fig. 2.15a, reproduce de mejor manera la tendencia lo cual implica que una mayor dimensión generará mayor inestabilidad del sistema.

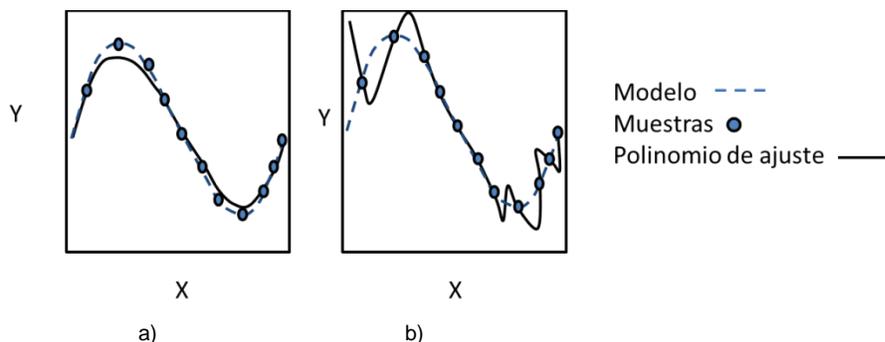


Fig. 2.15. Maldición de la dimensionalidad (Modificado de Bishop, 1995)

⁴ Característica propia de una señal (Bishop, 1995).

Por otro lado si el número de neuronas ocultas es menor que el número óptimo, la red es incapaz de aprender el mapeo entrada-salida correcto. Por lo tanto es importante determinar el número óptimo de neuronas ocultas para un problema dado (Alpana, 2002).

La generalización es influenciada por tres factores:

- i) El tamaño del conjunto de entrenamiento
- ii) La arquitectura de la red neuronal
- iii) La complejidad del problema a resolver.

No se puede tener control del último punto. En el contexto de los primeros dos, para obtener una buena generalización, se puede variar el tamaño del conjunto de patrones de entrenamiento (de entrada y de salida) manteniendo la arquitectura de la red fija o viceversa.

Este problema puede ser resuelto en términos de la dimensión de Vapnik-Chervonenkis (VC), la cual es una medida de la familia de funciones de clasificación realizadas por la red, es decir, de la capacidad de clasificación. La dimensión de Vapnik-Chervonenkis puede definirse como el número máximo de muestras de patrones de entrenamiento de entrada, para el cual una función puede clasificar correctamente todos los patrones en un conjunto de entrenamiento.

Los límites especificados por la dimensión VC puede ser declarados de la forma siguiente: "Si se desea un nivel de precisión de al menos 90%, entonces se debería usar diez veces el número de patrones de entrenamiento de entrada, como pesos haya en la red" (Alpana, 2002). Lo mismo ha sido sugerido por Poulton (2001). Baum y Hausler señalaron en 1989, que si se desea tener una precisión de $1 - \epsilon$, el número de muestras de entrenamiento debería ser aproximadamente $\frac{w}{\epsilon}$, donde ϵ es el error (en este caso en porcentaje entre la salida de la red y la salida deseada) y w el número de elementos del vector de pesos. En función de lo previamente explicado Alpana (2002), utilizó arquitecturas tales que el número de parámetros libres es más pequeño que el número de patrones de entrenamiento de entrada, diseñándolas mediante el uso de registros geofísicos de pozo sintéticos.

2.6.1 Dilema de la Varianza (error aleatorio) y Bias (error sistemático)

El error sistemático o bias (no confundir con el factor de polarización b) y la varianza pueden ser estimados cuando un predictor se entrena con diferentes conjuntos de datos muestreados aleatoriamente de un conjunto de entrenamiento (Alpana, 2002). En una red neuronal el término bias o error sistemático puede definirse como una medida de la habilidad que tiene el predictor de generalizar correctamente en una serie de pruebas y está relacionado con el grado de ajuste de los datos. Por otro lado la varianza de una red neuronal (llamada error aleatorio), se puede definir como una medida de hasta qué punto, la salida de una red es sensitiva a los datos con que fue entrenada, es decir, ¿Se obtendrían los mismos resultados si se hubieran utilizado otros datos del mismo conjunto de entrenamiento?

Ambos términos pueden estimarse de la relación del error cuadrático medio (Ec. 2.18) que existe entre una función objetivo (asociada con el término d de la Ec. 2.10) y el predictor (función que relaciona el conjunto de entrenamiento D definido por la Ec. 2.11 y los datos de entrada x); más detalles sobre esta relación pueden encontrarse en Geman, et. al. (1992).

$$(E_D[F(x; D)] - E[d|x])^2 + E_D[(F(x; D) - E_D[F(x; D)])^2] \quad 2.21$$

Término de error sistemático Término varianza

Donde E_D es el operador de esperanza, F es el estimador y $E[d|x]$ es la función objetivo.

Debe existir un balance entre el error sistemático y la varianza en términos del entrenamiento de las redes neuronales. Lo ideal sería contar con una varianza y un error sistemático pequeño. Sin embargo

reducir el error sistemático, probablemente resulte en un incremento en la varianza, mientras que para disminuir la varianza se suele dar lugar a un mayor error sistemático.

Esto conduce a una discusión acerca de dos conceptos:

- i) Sobre-entrenamiento.
- ii) La validación cruzada.

El sobre entrenamiento de una red comienza cuando se entrena hasta el límite (error deseado) o hasta que el número de epochs de entrenamiento es alcanzado. Si este proceso se repite varias veces para la misma entrada, siempre y cuando la configuración de la red lo permita, el resultado final será una red con error nulo para ese valor (La red memoriza)⁵. La red entrenada de esta forma generará un error sistemático más pequeño, pero de varianza más grande.

El enfoque de la validación cruzada, consiste en dividir el conjunto de datos de entrenamiento en dos subconjuntos (ajenos); el subconjunto de estimación y el subconjunto de validación. El subconjunto de estimación se utiliza para el entrenamiento. La sesión de entrenamiento se interrumpe periódicamente para probar la red en el subconjunto de validación y medir el error de validación. El entrenamiento de la red continúa siempre que el error cuadrático medio en el subconjunto de validación esté disminuyendo. El entrenamiento se detiene cuando el error de validación comienza a aumentar.

Después de que el ajuste de los pesos y los factores de polarización (bías) se detiene, el rendimiento de la red puede validarse con datos nuevos. El propósito del proceso de validación es determinar si la red es capaz de resolver el problema para la cual es entrenada. Este método se conoce como “early stopping method”.

La Fig. 2.16 muestra la variación del error con los dos tipos de metodologías de entrenamiento. El mismo método también ha sido sugerido por Poulton (2001) con el fin de saber cuándo detener el entrenamiento.

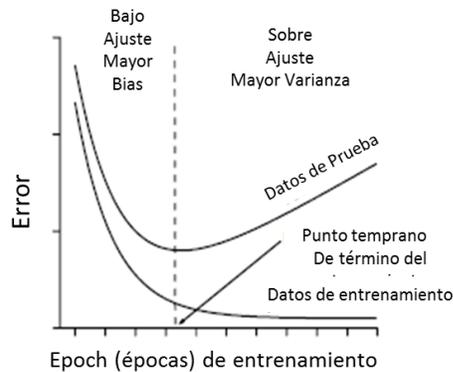


Fig. 2.16 Error y sobre entrenamiento.

⁵ Este no es el objetivo del entrenamiento, sino lo que se desea es entrenar a la red con varias entradas y luego ver qué sucede cuando ingresamos alguna que no estaba en el conjunto de datos de entrenamiento. Es por esta razón que no se itera sobre un mismo elemento del conjunto de datos hasta eliminar el error sino que se realiza un acercamiento con un elemento, luego con otro y así hasta recorrer todo el conjunto de datos. A esta recorrida sobre el conjunto de datos es otra definición de que es un epoch.

2.6.2 Ventajas y desventajas de una red MLP

Las principales ventajas de la red MLP son (Alpana, 2002):

- a) Las redes *MPL* (y en general de todos los modelos de redes neuronales) generan una relación no lineal entre las entradas y salidas.
- b) El conocimiento previo de la relación entre la entrada y salida no es necesario a diferencia de los métodos estadísticos. Alpana (2002) menciona que no se requiere de ninguna suposición sobre la distribución subyacente de los datos para el diseño de las redes; sin embargo, en este trabajo proponemos el análisis estadístico como una herramienta para mejorar el rendimiento de la red neuronal durante entrenamiento (ver capítulo 3).
- c) Su estructura paralela hace que sea fácilmente realizable en computadoras trabajando en paralelo.
- d) La red es robusta y presenta un alto grado tolerancia a fallos debido a la redundancia intrínseca.
- e) El daño de unos pocos nodos o enlaces no perjudica el rendimiento global de forma significativa.
- f) Se puede formar cualquier región de decisión, en el espacio generado por las entradas sin ninguna restricción. Estas regiones incluyen polígonos convexos y regiones convexas.
- g) La red tiene una gran capacidad para la aproximación de funciones. Las habilidades para aprender y generalizar, son cualidades adicionales.
- h) La MLP tiene la capacidad intrínseca de adaptar sus pesos sinápticos a cambios en el entorno que lo rodea al ajustar los pesos para minimizar el error.

Dentro de las desventajas se enumeran las siguientes (Alpana, 2002):

- a) Para una red MLP la topología es importante para la solución de un determinado problema, es decir, definir el número de neuronas ocultas y el tamaño los datos de entrenamiento, lo cual es generalmente la parte más difícil de definir.
- b) A causa de la no linealidad y la compleja conectividad que presenta la red, es difícil realizar un análisis teórico de la respuesta de la red. El uso de neuronas ocultas hace que el proceso de aprendizaje sea más difícil de visualizar. Alpana (2002) propone el uso de datos sintéticos y la teoría de la dimensión VC, para identificar la arquitectura óptima de la red. En términos generales para mejorar el rendimiento del proceso de entrenamiento, en este trabajo se utilizan datos sintéticos generados de relaciones petrofísicas y datos estadísticos de campos análogos (Ver capítulo 4).
- c) Debido a la alta relación no lineal, se tienen tasas de entrenamiento muy bajas en las redes MLP muy grandes, cuando se usa el algoritmo de retropropagación, lo que se traduce en procesos de convergencia muy lentos. Muchas veces se utiliza el algoritmo Levenberg-Marquardt ⁶ para superar este problema.
- d) La superficie del error cuadrático medio de una red MLP puede tener muchos mínimos locales y un mínimo global, existiendo la posibilidad de quedar atorado en los mínimos locales en vez de converger hacia el mínimo global.

⁶ Más detalles sobre este algoritmo pueden verse en More (1978).

- e) La salida de una sola red puede no ser satisfactoria. Con el fin de evitar este problema se utilizan redes múltiples en lugar de una sola red de modo que si una de las redes da un mayor error, la contribución de ésta red es reducida en la salida del conjunto; esto es conocido como el enfoque de prueba y selección (Sharkey et al., 2000).

2.7 Mapas auto-organizados

La idea del aprendizaje competitivo, se vislumbró en los primeros trabajos de Von der Malsburg (1973) sobre la auto-organización de las células nerviosas de la corteza cerebral. En 1975, Fukushima propuso el cognitrón que es una red competitiva multicapa y auto-organizada. Willshaw y Von der Malsburg (1976) trabajaron sobre la formación de las conexiones neuronales mediante auto-organización y Grossberg (1972, 1976) trabajó sobre la clasificación adaptativa de patrones.

Rumelhart y Zisper especificaron en 1985 los tres elementos básicos de una regla de aprendizaje competitiva:

- 1- Un conjunto de neuronas (unidades de proceso) que se activan o no en respuesta a un conjunto de patrones de entrada (estímulos) y cada una de ellas tienen asignado un valor, entre un conjunto de pesos sinápticos, por lo que responderán de manera diferente a un mismo estímulo.
- 2- Un límite impuesto sobre los pesos sinápticos de cada neurona, para que se activen.
- 3- Un mecanismo que permite competir a las neuronas para responder a un subconjunto de entradas, de tal manera que una y sólo una neurona por grupo se activa.

Los mapas auto-organizados o también llamados mapas de preservación de la topología, asumen que entre grupos y sus unidades (neuronas) existen estructuras topológicas definidas. Esta propiedad se observa en el cerebro, pero no se observa en otros tipos de redes neuronales (Fausset, 1994). El principal objetivo de los mapas auto-organizados es preservar la topología de los datos multidimensionales en un espacio de menor dimensión de tal forma que se preserve su estructura y sus propiedades (se preserven los grupos, las relaciones entre ellos, etc.) (Stefanovic y Kurasova, 2011).

Los mapas auto-organizados son un conjunto de neuronas conectadas entre sí, mediante pesos sinápticos, que forman topologías rectangulares o hexagonales (figura 1.17), en donde la salida representa grupos en los datos de entrada. En este tipo de redes los datos se proyectan a un plano (Generalmente bidimensional).

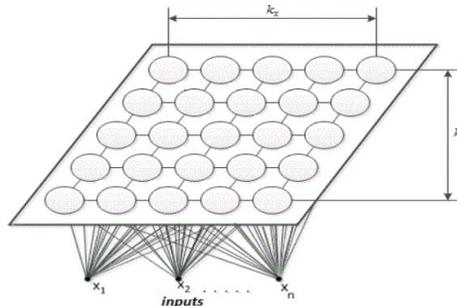


Fig. 2.16 Arquitectura rectangular de un mapa auto-organizado

El conjunto de pesos forma un vector W_{ij} , con $i=1, \dots, k_x$ y $j=1, \dots, k_y$, en donde k_x es el número de filas y k_y el número de columnas en un mapa auto-organizado con topología rectangular. La dimensión del vector de pesos es igual que el número de entradas x^n , donde $W_{ij} = \{w_{ij}^1, w_{ij}^2, \dots, w_{ij}^n\}$ (Stefanovic y Kurasova, 2011).

Los mapas auto-organizados constan solo de dos capas, la capa de entrada y de salida (no consta de capas ocultas) y su aprendizaje es de tipo no supervisado, el cual, comienza con la inicialización de los componentes del vector de pesos W_{ij} , de forma aleatoria. Si la entrada a una red es conformada por los

vectores n-dimensionales $X_p \in \{X_1, X_2, \dots, X_m\}$ y cada vector X_m , tiene componentes x_1, x_2, \dots, x_n , en cada paso del aprendizaje, un vector X_p se compara con los pesos de W_{ij} de todas las neuronas.

Usualmente la forma de compararlo es con la distancia Euclidiana entre el vector de entrada y el peso calculado $\|X_p - w_{ij}\|$. La neurona con la menor distancia a X_p es la ganadora (Stefanovic y O. Kurasova, 2011).

En la fase de aprendizaje, cada neurona del mapa sintoniza (o mapea) con diferentes rasgos en el espacio de entrada. Tras la presentación y procesamiento de un vector de entradas $X_p(t)$ la neurona vencedora modifica sus pesos de manera que se parezcan un poco más a $X_p(t)$. De este modo, ante el mismo patrón de entrada, dicha neurona responderá con más intensidad en el futuro (Aguilar, 2008).

El proceso se repite para numerosos patrones de entrada, de forma que al final los diferentes vectores de pesos sintonicen con dominios específicos de las variables de entrada y tiendan a representar la función de densidad de probabilidad $p(x)$ del espacio de entradas (Aguilar, 2008). Si dicho espacio está dividido en grupos, cada neurona se especializará en uno de ellos y la operación esencial de la red se podrá interpretar entonces como un análisis de agrupamiento (Aguilar, 2008).

De manera simplificada los pesos de las neuronas son adaptados con la siguiente regla de aprendizaje (Mehotra et al., 1997):

$$w_{ij}(t + 1) = w_{ij}(t) + \eta(t)(X_p - w_{ij}(t)) \quad 2.22$$

Donde t es el número de iteración y η es la tasa de aprendizaje.

La tasa de aprendizaje es una función monótonamente decreciente con el tiempo, siendo habitual su actualización mediante una función lineal:

$$\eta(t) = \eta_0 + (\eta_f - \eta_0) \frac{t}{t_f} \quad 2.23$$

Donde η_0 es la tasa de aprendizaje inicial, η_f es la tasa de aprendizaje con la que se desea terminar y t_f , es el número de iteraciones con el que se espera alcanza el valor η_f .

Otra aproximación es:

$$\eta(t) = \eta_0 \left(1 - \frac{t}{t_f}\right) \quad 2.24$$

Una alternativa es usar una función que decrezca exponencialmente:

$$\eta(t) = \eta_0 \left(\frac{\eta_f}{\eta_0}\right)^{\frac{t}{t_f}} \quad 2.25$$

También suele considerarse como una elección razonable

$$\eta(t) = 0.9 \left(1 - \frac{t}{1000}\right) \quad 2.26$$

El empleo de una u otra función no suele influir demasiado en el resultado final. No siendo así en el caso de mapas muy grandes, donde la selección de un valor óptimo es crucial para la convergencia (Aguilar, 2008).

Lo descrito hasta el momento responde a un esquema competitivo clásico de relativa sencillez, en el que cada neurona actúa en solitario. Sin embargo, el modelo de SOM aporta una novedad importante, pues incorpora a este esquema relaciones entre las neuronas que se encuentran cercanas a la neurona ganadora.

En 1982 el profesor de la Universidad de Helsinki Teuvo Kohonen, presentó un modelo sencillo para la formación auto-organizada de mapas de características de los datos de entrada, explotando la cooperación entre las neuronas vecinas.

Este modelo incluye el término de una función de vecindad $h(i-g,t)$, que determina, si cada neurona i , pertenece o no a la vecindad de la neurona vencedora g . Esta función decrece con la distancia a la neurona ganadora y dependerá de un parámetro denominado radio de vecindad $R(t)$ que representa el tamaño de la vecindad actual. La función $h(i-g, t)$, representa el complejo sistema de interacciones laterales entre neuronas (Aguilar, 2008). La función de vecindad más simple es del tipo escalón:

$$h(\|i - g\|, t) = \begin{cases} 0 & \text{si } \|i - g\| > R_t \\ 1 & \text{si } \|i - g\| \leq R_t \end{cases} \quad 2.27$$

Con este tipo de función, las vecindades adquieren forma circular, de bordes nítidos, en torno a la neurona vencedora y la Ec. 2.22 puede escribirse y reducirse a:

$$\Delta_{w_{ij}} = w_{ij}(t + 1) = w_{ij}(t) + \eta(t)h(\|i - g\|, t)(X_p - w_{ij}(t)) \quad 2.28$$

$$\Delta_{w_{ij}} = \begin{cases} 0 & \text{si } \|i - g\| > R_t \\ w_{ij}(t) + \eta(t)(X_p - w_{ij}(t)) & \text{si } \|i - g\| \leq R_t \end{cases} \quad 2.29$$

Por lo que en cada iteración únicamente se actualizan las neuronas que distan de la vencedora en una distancia menor a $R(t)$.

La función de vecindad posee una forma definida, pero su radio $R(t)$ varía con el tiempo. Se parte de un valor inicial R_0 grande, (incluso puede ser más de la mitad del diámetro de la red), que determina vecindades amplias, con el fin de lograr la ordenación global del mapa. $R(t)$ decrece monótonamente con el tiempo, hasta alcanzar un valor final de $R_f=1$ con el que solamente se actualizan los pesos de la neurona vencedora y las adyacentes. Aunque existen muchas funciones de actualización, una posible función es:

$$R(t) = R_0 + (R_f - R_0) \frac{t}{t_f} \quad 2.30$$

Donde R_0 es el radio de vecindad inicial, R_f es el radio de vecindad final propuesto, t es la iteración y t_f el número de iteraciones propuesto para alcanzar R_f . Si únicamente se dispone de un número relativamente pequeño de muestras, entonces éstas deberán ser recicladas para alcanzar el número deseado de iteraciones.

Dentro del modelo de Kohonen se puede utilizar una medida de similitud (correlación) entre los pesos y las entradas. Sin embargo, esta medida es sensible al tamaño de los vectores; grandes diferencias en sus longitudes pueden introducir una importante distorsión en la medida de similitud. Para resolver este problema puede dividirse por las normas de los vectores, con lo que se tiene el denominado criterio del coseno, aunque existen otras como la distancia de Minkowski y la de Manhattan u otras relacionadas con la correlación, como la medida de similitud de Tanimoto, la de Mahalanobis o la de Hamming,

La regla de actualización se obtiene a partir de la optimización de una función objetivo o de error E (definida por las distancias euclidianas o correlaciones), a partir del algoritmo de descenso por gradiente. Es decir, habrá que modificar los pesos en la dirección opuesta al gradiente, esto es:

$$\Delta_{w_{ij}} = \eta \frac{dE}{dw_{ij}} \quad 2.31$$

Uno de los algoritmos de optimización de descenso por el gradiente más conocidos es el algoritmo de backpropagation, que modifica los valores de los parámetros proporcionalmente al gradiente de la función de error con objeto de alcanzar un mínimo local.

El objetivo será encontrar una función de error que permita deducir sistemáticamente reglas de aprendizaje. Como se pretende que los pesos ajusten la distribución de las entradas, un objetivo puede ser que los pesos sinápticos tiendan a ellas, es decir, que los errores de cuantificación sean lo más pequeños posibles. Con esta premisa, puede definirse una función objetivo global para la red de la manera siguiente.

$$E = \int \sum_k h(i - g) f(\|X_p - w_k\|) p(x) dx \quad 2.32$$

Dónde:

- 1- $h(i-g)$ función de vecindad
- 2- $f(\|X_p - w_i\|)$ es la función de medida de distancia o similitud
- 3- $p(x)$, es la función de probabilidad del espacio de entradas.
- 4- X_p , son los vectores de entrada y w_k son los pesos del vector w_{ij}

Esta función objetivo global (Ec. 2.32) se basa en la suma de todos los errores cuantificados en las neuronas, la cual esta ponderada en la vecindad y promediada por medio de la función de distribución de probabilidad para todas las entradas posibles.

Una aproximación obtenida por medio del método de descenso por gradiente es:

$$E_k(t) = \sum_k h(i - g, t) f(\|X_p - w_k\|) p(x) dx \quad 2.33$$

Cuya solución es:

$$w_k(t + 1) = w_k(t) + \eta(t) \nabla_{w_k} E_k(t) \quad 2.34$$

Donde la tasa de aprendizaje debe cumplir:

$$\sum_{t=0}^{\infty} \eta(t) = \infty \quad 2.35$$

$$\sum_{t=0}^{\infty} \eta^2(t) < \infty \quad 2.36$$

Al estar realizando descensos por los gradientes locales proporcionados por la función de error $E_k(t)$ (Ec. 2.33) y los no descensos por la máxima pendiente que proporciona la función E (Ec. 2.30), se obtienen soluciones no globalmente óptimas. No obstante, se ha mostrado que el punto que se alcanza está muy próximo al óptimo, y que puede considerarse que las soluciones que proporciona son casi óptimas. Este procedimiento permite deducir sistemáticamente algoritmos de aprendizaje sólo con cambiar el criterio de distancia $\|X_p - w_k\|$ y la función asociada a ella $f(\|X_p - w_k\|)$ (Aguilar, 2008).

En resumen, el proceso de entrenamiento de la red puede definirse en cuatro pasos:

- 1- Inicialización.- Todos los pesos son inicializados con pequeños valores aleatorios.

- 2- Competición- Para cada patrón de entrada, las neuronas calculan sus respectivos valores de una función discriminante la cual da la base de la competición. La neurona con menor valor de la función discriminante es declarada ganadora.
- 3- Cooperación.- La neurona ganadora determina la localización espacial de la vecindad topológica de neuronas excitadas y por lo tanto da la base para la cooperación entre las neuronas de la vecindad.
- 4- Adaptación. Los valores individuales de la función discriminante de las neuronas excitadas decrecen en relación al patrón de entrada a través de un adecuado ajuste de los pesos asociados, tal que la respuesta de la neurona ganadora es mejorada en la subsecuente aplicación de patrones de entrada que sea similar.

La red neuronal de Kohonen difiere considerablemente de la red neuronal de alimentación hacia adelante (feedforward) o de retropropagación (Backpropagation), tanto en la forma en que se entrena, cómo en la forma en que es recordado un patrón. La red neural Kohonen no utiliza ningún tipo de función de activación y no considera el vector de valores bias (Shatil, 2006). La red de Kohonen requiere que las entradas estén normalizadas entre -1 y 1 para su funcionamiento.

2.8 Sistema de múltiples redes

Un sistema de múltiples redes es una combinación de varias redes neuronales. Hay varios problemas que no pueden ser resueltos por una sola red o que puede ser resueltos más eficientemente por sistemas de múltiples redes (Alpana, 2002). Se puede lograr un mejor rendimiento, combinando redes de forma redundante en lugar de una sola red. Hay dos métodos comunes de crear un sistema de redes múltiples: La **combinación de conjunto (ensemble combination)** y la **combinación modular (modular combination)**.

Bajo un enfoque modular la tarea o el problema se descompone en una serie de sub-tareas y la solución completa requiere la contribución de todos los módulos encargados de realizarlas. Por el contrario, en un combinación de conjunto, las redes componentes son de tipo redundantes, ya que cada una de ellas proporciona una solución a la misma tarea o sub tarea, aunque esta solución se puede obtener por diferentes medios. Tanto las combinaciones en **conjunto** y **modulares** pueden existir en una tarea o una sub-tarea.

Las combinaciones de **conjunto** y **modulares** no son mutuamente excluyentes y un sistema real de múltiples redes podría consistir en una mezcla de combinaciones de **conjunto** y **modulares** en diferentes niveles. Haykin (1999) nombra a un sistema de múltiples redes como una máquina comité (Committee Machine – *CM*). Alpana (2002) utiliza un sistema de redes combinándolas con ambos enfoques (de conjunto y modulares) para estimar litofacies en el pozos del campo Oseberg en el Mar del Norte, mediante registros de geofísicos de pozo de herramientas LWD. En este trabajo de tesis se define el mismo enfoque.

2.8.1 Combinación de conjunto

La motivación principal para la combinación de redes en conjuntos es el de mejorar su capacidad de generalización o para prevenir el fracaso de los componentes individuales (Alpana, 2002). La razón para esperar que algunas veces las redes individuales puedan fallar o cometer errores con algunas entradas, se basa en la suposición de que han sido entrenadas con un conjunto datos de entrenamiento limitado y que estos son necesarios para estimar la función objetivo. Tales estimaciones no serán idénticas a la función objetivo a menos que el conjunto de datos sea infinitamente grande. El error cuadrático medio de un predictor se expresa en términos del error sistemático (bias) y la varianza de la red (Ec.. 2.21).

La expectativa de generar una **combinación de conjunto** es que los diferentes expertos de entrenamiento converjan a diferentes mínimos locales en la superficie de error y en el rendimiento global se mejoré de alguna manera, mediante la combinación de sus diferentes salidas.

De un análisis de errores presentado en Naftaly et al (1997) para este tipo de combinación de redes se encontró que:

- i) El bias (error sistemático) de una función promedio del conjunto de redes $F_1(x)$ es exactamente la misma que la función $F(x)$ perteneciente a una simple red neuronal.
- ii) La varianza de la función promedio del conjunto $F_1(x)$ es menor que la de la función $F(x)$.

Lo anterior apunta a una estrategia de entrenamiento para reducir el error total producida por el sistema múltiple de redes debido a la variación de las condiciones iniciales (es decir, los pesos iniciales aleatorios). El experto individual debería ser deliberadamente sobreentrenado para reducir el error sistemático (bias) a costa de la varianza (Alpana, 2002).

El objetivo de este enfoque es encontrar redes, que generalicen de manera diferente, ya que pueden existir redes que podrían variar en términos de sus pesos, en términos del tiempo que tarda en converger (parámetros del algoritmo) e incluso su arquitectura, sin embargo constituyen esencialmente la misma solución, ya que resultan en el mismo patrón de errores durante la etapa de prueba. Existe un número de parámetros de entrenamiento, que pueden ser manipulados con este objetivo en mente. Estos incluyen los siguientes: los pesos iniciales, los datos de entrenamiento, la topología de las redes y el algoritmo de entrenamiento. Bajo este esquema se pueden mantener constantes algunos parámetros y variar otros, en función del problema en estudio; por ejemplo, en Alpana (2002) se generó una combinación de redes de **conjunto** para predecir porosidad y saturación de agua, mediante la variación aleatoria de los pesos iniciales, manteniendo los datos de entrenamiento constantes; mientras que para estimar litofacies se varían los datos de entrenamiento y se respeta la misma topología.

Los datos de entrenamiento pueden ser variados por boosting (Schapire, 1990), donde una serie de aprendices débiles (en este caso las redes individuales), pueden convertirse en un solo aprendiz (el conjunto de redes neuronales), mediante el entrenamiento de cada red individual con patrones que han sido procesados por otras redes individuales (del mismo conjunto). El problema de este método es que requiere una gran cantidad de datos de entrenamiento. Freund y Schapire (1996), propusieron el algoritmo AdaBoost que elimina este problema, generando un remuestreo adaptativo, de tal modo que los pesos son incrementados durante el remuestreo, para los casos en donde existe una mala clasificación. Otro método para variar los datos de entrenamiento es por bootstrapping (Breiman, 1996), en donde se crean distintas versiones de los datos de entrenamiento, tomando n muestras a la vez de un conjunto de datos de entrenamiento de N muestras por sustitución. Este procedimiento puede repetirse varias veces para crear un número de conjuntos de datos diferentes y traslapados. Si existe escasez de patrones de entrenamiento, como normalmente es el caso en datos reales, entonces, muestras virtuales pueden ser generadas como se sugiere en Cho et al. (1997). Un método similar es el uso de sets de entrenamiento mutuamente excluyentes; es decir, con muestreo sin reemplazo (Sharkey et al, 1996.). Entonces no hay traslapes entre los datos, utilizados para entrenar las diferentes redes y la diversidad en cuanto a la generalización se garantiza.

En Alpana (2002), se genera un muestreo sin reemplazo para las redes con enfoque de conjunto (redes neuronales de conjunto o ensemble neural network) y el muestreo con reemplazo cuando se trabaja con los diferentes módulos (enfoque modular, redes neuronales modulares o modular neural network)

Una vez que se ha creado un conjunto de redes, se debe encontrar una forma eficaz de la combinar las diferentes salidas. Hay varios métodos diferentes para combinar las redes dependiendo de la tarea realizada por las redes es decir si se está ejecutando aproximación de funciones o de reconocimiento de patrones. Una salida simple se puede crear a partir de un conjunto de salidas de redes, mediante el promedio del conjunto (Naftaly et al., 1997) o por medio de un promedio ponderado que tome en cuenta

las precisiones relativas de las redes que se combinan, sugiriéndose dos métodos: sin restricciones y limitada (Hashem, 1997).

En el enfoque no restringido, para asignar diferentes pesos a la salida de una red individual, la salida combinada de los K expertos puede ser expresada por la suma ponderada:

$$\tilde{y}(x, \alpha) = \sum_{k=1}^K \alpha_k y_k(x) \quad 2.37$$

Donde y_k es la salida de las redes individuales α_k , $k=1..K$, son los pesos y $x = \{x_i; 1 = 1 \dots N\}$ es el vector de entrada. El requisito es evaluar α_k de forma que y_k proporcione una estimación de mínimos cuadrados de la respuesta deseada $d(x)$ (d en la Ec. 2.10). Dado un conjunto de entrenamiento el error de aproximación es:

$$\delta(x; \alpha) = d(x) - \tilde{y}(x; \alpha) \quad 2.38$$

Hashem (1997), incluyó el término α_0 , para extender la definición de $\tilde{y}(x; \alpha)$

$$\tilde{y}(x, \alpha) = \sum_{k=0}^K \alpha_k y_k(x) \quad 2.39$$

En su trabajo Hashem (1997), demostró que la combinación lineal óptima (OLC) de los pesos dados sin restricción es equivalente a una regresión de $d(x)$ y $y_k(x)$, con $k = 1, \dots, K$, donde α_0 es la ordenada al origen.

En la aproximación con restricción hay una restricción adicional que es:

$$\sum_{k=0}^K \alpha_k = 1 \quad 2.40$$

2.8.2 Red Modular (Modular Neural Network)

La descomposición modular puede llevarse a cabo con el propósito de mejorar el rendimiento. Una tarea puede ser resuelta con una simple red, pero mejorar el rendimiento se logra cuando una tarea se desglosa en una serie de subtareas realizadas por módulos especialistas. El enfoque de divide y vencerás (Jordan y Jacobs, 1994) que es ejemplificado por la mezcla de aproximaciones de expertos (Haykin, 1999) ofrece un ejemplo de la mejora del rendimiento que puede resultar de un sistema modular. Si una sola red reemplaza la combinación de varios expertos, esta red tendrá un número grande de parámetros ajustables y por lo tanto el riesgo de sobreajustar del conjunto de entrenamiento aumenta (Alpana, 2002). Otra razón para adoptar un enfoque modular es el de reducir la complejidad del modelo, y hacer que el sistema global sea más fácil de entender, modificar y extender. Los tiempos de entrenamiento pueden reducirse como resultado de la descomposición modular, ya que probablemente el tiempo de entrenamiento de una red muy grande probablemente será más largo que para todos los expertos entrenados en paralelo. Cada subproblema podría solucionarse con una diferente arquitectura de red neuronal o un algoritmo que permita explotar resolver cada caso específico. El conocimiento previo puede ser incorporado en términos de sugerir una descomposición adecuada de una tarea. El pre-procesamiento de las entradas de una red neural antes del entrenamiento se considera también como una forma de descomposición modular con el fin de simplificar el problema en cuestión.

El factor más importante de un sistema modular es la forma en que los componentes modulares han sido creados, ya sea para resolver una tarea específica o bien para combinar diferentes entradas en una salida única. Una tarea puede descomponerse en subtareas, cada una de las cuales es resuelta por un módulo de la red. El método propuesto por Anand et. al., (1995) consiste en dividir un problema de clasificación de k-clases en un problema de clasificación con k conjuntos de clases binarias (0 y 1), mientras se utiliza el mismo número de datos de entrenamiento que el problema original de k clases. Este

principio es muy socorrido en la predicción litofacies, donde cada componente modular es una red neuronal entrenada para una simple litofacie con clase binaria (1 pertenece a la clase, 0 no pertenece a la clase). La descomposición de tareas se puede utilizar para realizar una tarea la cual no se podría lograr mediante una sola red o no podría ser llevada a cabo con la misma facilidad o de forma natural.

La mezcla de expertos y mezclas jerárquicas de expertos dividen los datos en regiones y ajusta superficies simples a los datos que se encuentran en cada región del espacio de los datos de entrada. Redes de expertos aprenden a especializarse en subtareas y cooperar a través de una red global. Los límites de las regiones no son rigurosos, lo que significa que los puntos de los datos pueden estar simultáneamente en varias regiones (los límites son difusos).

Hay cuatro modos diferentes de combinar los componentes de la red modular tal como se ven en Sharkey (1999), cooperativo, competitivo, secuencial y de supervisión. La principal diferencia entre las combinaciones cooperativa y competitiva es que, en la primera se asume que todos los elementos que se combinan harán alguna contribución a la decisión, aunque esta contribución puede ser ponderada de alguna manera, mientras que en la combinación competitiva se asume que para cada entrada se selecciona el elemento más adecuado, en la combinación secuencial el procesamiento es sucesivo, en donde el cálculo de un módulo depende de la salida del módulo anterior y en la combinación de supervisión, un módulo se usado para supervisar el desempeño de otro módulo.

2.9 Regresión Lineal Multidimensional.

Al igual que las redes neuronales, el análisis de regresión también se refiere a la formulación de modelos matemáticos que describen las relaciones entre las variables y al uso de estas relaciones modeladas con el objetivo de predecir y realizar otras inferencias estadísticas (Alpana,2002). En muchos casos la relación entre las variables no es determinista, pero las variables están mutuamente relacionadas, es decir una variable depende de una serie variables causales donde la relación no se rige por una ley física precisa, por ejemplo, una gráfica de valores observados de variables describe una relación entre ellas.

Después de obtener suficiente conocimiento acerca de una relación empírica puede ser posible para el investigador formular una teoría que conduzca a una fórmula matemática y por lo tanto el caso semi-determinista.

En la regresión lineal multidimensional, los hiperplanos que definen la regresión de las diferentes variables, están en función de los llamados parámetros de regresión. El modelo de regresión sugiere que existe dependencia lineal con cada una de las variables independientes cuando las otros dos se mantienen fijas. Los parámetros de regresión se estiman por el principio de mínimos cuadrados.

2.9.1 Analogía de la regresión lineal Multidimensional con las redes neuronales.

El método para identificar la relación entre una variable respuesta y varias variables causales usando Regresión Lineal Multidimensional es análogo al procedimiento de identificar las relaciones de entrada y salida de las variables usando redes neuronales. Los parámetros de regresión son similares a los pesos de la red y el bias (factor de polarización). Los parámetros de la regresión son estimados por el principio de mínimos cuadrados, minimizando la suma del cuadrado de las desviaciones estándar, similarmente a los pesos y los bias de una red neuronal que son ajustados minimizando una función de costo.

La diferencia entre las redes neuronales y la regresión multilineal es que la relación de las entradas y salidas es lineal en la regresión lineal multidimensional, mientras en las redes neuronales no lo es (Alpana, 2002). El método de redes neuronales a diferencia del método de regresión lineal multidimensional no necesita forzar los valores a yacer cerca de los valores medios y así preserva la variabilidad natural en los datos.

III. METODOLOGÍA

El proceso común para la estimación de clasificación de facies se basa en una interpretación manual, sin embargo el uso de técnicas estadísticas como el análisis de componentes principales o análisis discriminantes, para posteriormente utilizar algún algoritmo clasificador (Análisis de agrupamiento o redes neuronales) es una práctica común en la industria (Casar et. al., 2012).

Como se explicó anteriormente el principal objetivo de este trabajo de tesis es el planteamiento y análisis de un proceso metodológico para la estimación de electrofacies a partir de la clasificación de registros geofísicos de pozo, mediante el uso de modelos neuronales de aprendizajes supervisados y no supervisados.

Como se planteó en el capítulo anterior, el enfoque natural que siguen los modelos neuronales, es el estadístico, por lo que la metodología propuesta, sugiere que el conocimiento previo del comportamiento y naturaleza estadística de los datos es un paso fundamental durante la etapa de entrenamiento de una red neuronal, teniendo una repercusión directa durante el proceso de clasificación de los datos y por ende de los resultados.

No obstante la importancia que tiene esta información, en muchos trabajos metodológicos donde las redes neuronales juegan un papel protagónico, se omite o se da por entendida la naturaleza de los datos, lo cual conlleva a realizar ensayos de prueba y error, de manera indiscriminada. Durante cualquier proceso de investigación (aplicación de un modelo) los ensayos de prueba y error son naturales e inevitables, sin embargo, el conocimiento que podamos obtener de nuestros datos, permite realizarlos de forma más metodológica, de forma ordenada y progresiva, con los criterios suficientes que nos permitan tener un mejor control de dichas pruebas.

No menos importante es la información obtenida de los resultados y los errores, que se obtienen no solo durante la aplicación del modelo mismo, sino también los obtenidos durante las etapas de entrenamiento y validación de una red neuronal. Estos análisis, permiten entender la evolución de nuestro modelo neuronal y con ello optimizar y validar nuestros resultados.

Debido a que dos de modelos neuronales (*MPL* simple y *CM*), que se proponen en este trabajo, son de aprendizaje supervisado, es necesario contar con una clasificación previa de facies o clases petrofísicas (maestro), interpretadas en los registros geofísicos realizadas por algún experto en la materia (ver capítulo anterior y capítulo 5), sin embargo, es conveniente realizar un control de calidad a dicha clasificación, en función del conocimiento del área, los datos adicionales que se puedan tener (núcleos, análisis granulométricos, sísmica, etc.) y la coherencia entre ellos; ya que una práctica común es la edición de los registros de pozo realizadas por un petrofísico experto, como lo son: la normalización de registros mediante histogramas (Fig. 3.1), pegado y traslape de curvas (durante la edición de diferentes corridas), ediciones en profundidad y/o estimación de curvas en intervalos faltantes (o curvas completas). Debido a lo anterior los datos se ven muchas veces afectados e influenciados por el petrofísico especialista, generándose situaciones como evaluaciones con diferentes criterios (debido a que pueden ser diferentes especialistas en diferentes tiempos, diferentes criterio dependiendo de la posición de los pozos, etc.)

La consistencia de la clasificación juega un papel muy importante durante la aplicación de una red neuronal supervisada. En este trabajo de tesis proponemos como hipótesis de trabajo, que los resultados de los modelos neuronales (supervisados y no supervisados) así como las técnicas de clústeres, permiten identificar inconsistencias en la clasificación.

Las redes neuronales plantean algoritmos poderosos para la clasificación de datos y estimación de funciones, sin embargo el éxito de éstos, está en función de los datos con que son entrenados y la validación de los mismos, es por ello que la propuesta metodológica plantea en primer instancia un flujo de trabajo, cuyo fundamento se centra principalmente en el análisis exploratorio de datos y la selección

de variables durante la etapa de entrenamiento, así como en la validación cruzada y el análisis del error, en las etapas de prueba y aplicación de la red.

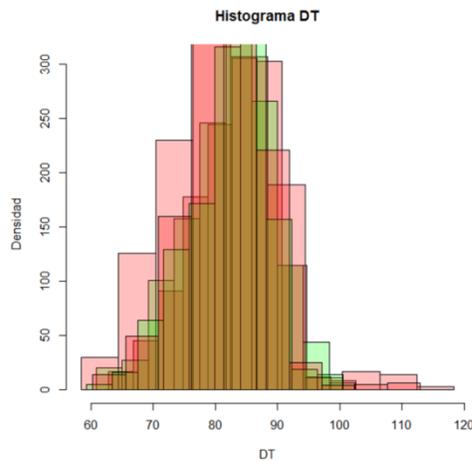


Fig. 3.1 Ejemplo de Histogramas normalizados por un petrofísico experto.



Fig. 3.2 Flujo trabajo propuesto

El flujo de trabajo propuesto consta de los pasos siguientes:

- 1) Censo y revisión de la información (control de calidad).
- 2) Análisis Exploratorio de Datos (AED).
- 3) Análisis de agrupamiento (Análisis de clústeres)
- 4) Análisis de selección de variables (Análisis de componentes principales)
- 5) Diseño y entrenamiento de la red neuronal.

- 6) Análisis del error y validación cruzada
- 7) Clasificación multi-pozo de electrofacies mediante el modelo neuronal.

El flujo de trabajo se resume en la Fig. 3.2

3.1 Censo y revisión de la información (control de calidad)

Esta etapa de la metodología en ocasiones se considera obvia y aunque es la base fundamental de todo trabajo, muchas veces se omite, ya sea por el desconocimiento de la disciplina en estudio, porque confiamos en la fuente de información o porque la experiencia en la materia permite inferir la calidad de los datos.

El conocimiento, ordenamiento y clasificación de la información con la que contamos, así como los controles de calidad que podamos generar, permite realizar un primer análisis de factibilidad, sobre si el proceso que se aplicará puede ser realizado o no. Para nuestro problema en cuestión, el conocimiento sobre geología, petrofísica y análisis de registros geofísicos de pozo, es deseable, ya que permitirá tener criterios de discriminación a la hora de categorizar y organizar la información.

El primer paso dentro la metodología propuesta, es conocer con qué tipo de registros geofísicos se cuenta para cada pozo, ya que de esto depende, seleccionar los pozos que serán utilizados durante el entrenamiento de la red neuronal. Es recomendable ordenarlos en función de sus características (propiedad física, respuesta al principio físico, p. ej., si el principio físico de la herramienta responde a litologías, a la matriz, porosidad y al contenido o tipo de fluidos), en función de su utilidad para establecer electrofacies y en función de la equivalencia entre curvas (Análisis de mnemónicos), en el caso de que los pozos en estudio hayan sido tomados con diferentes herramientas, en condiciones distintas o hayan sido corregidos por diferentes causas (ya que se requiere contar con datos equivalentes para estimar las electrofacies).

El control de calidad de los registros geofísicos de pozo es principalmente parte del trabajo del analista de registros geofísicos de pozo y del petrofísico, el cual está en función del dominio de estas disciplinas y el conocimiento geológico del área. Sin embargo, siempre es recomendable analizar las condiciones en que se encuentran los datos, ya que muchas veces los archivos, pueden estar corruptos, con datos faltantes, valores anómalos, ruidosos, datos que necesitan ser transformados (herramientas calibradas bajo ciertas condiciones como en matriz de caliza o arenisca, p. ej., el registro neutrón). Es necesario verificar si entre las curvas de un pozo, no se presentan desplazamientos en profundidad y sería altamente recomendable analizar las condiciones en que fueron tomados (diseño y condiciones de perforación), revisar las etapas de perforación y las corridas para identificar efectos anómalos (p. ej. Efectos causados por asentamientos de tuberías de revestimiento, problemas en la edición de registros entre corridas o si los datos fueron probablemente estimados). Lo anteriormente planteado, puede ser identificado mediante el apoyo de las técnicas empleadas en la siguiente etapa de la metodología (análisis exploratorio de datos).

3.2 Análisis exploratorio de datos (AED).

La segunda etapa de esta metodología, se conoce como análisis exploratorio de datos, la cual está basada en técnicas estadísticas convencionales y gráficas que permiten obtener todo un conjunto de información, establecer un buen entendimiento básico del comportamiento de los datos y de las relaciones existentes entre las variables que se estudian (Díaz, 2004).

El análisis exploratorio de datos (AED) fue introducido por Jhon W. Tukey en 1977, como un conjunto de técnicas para encontrar patrones de interés en los datos y lo robusto de los métodos no paramétricos, los cuales hacen pocas suposiciones sobre la forma de la distribución de los datos. El AED utiliza la

estadística univariada, multivariada, los análisis de regresión y de mínimos cuadrados, como herramientas (Díaz, 2004).

El análisis exploratorio de datos (*AED*) es un paso previo e indispensable para la aplicación exitosa de cualquier método estadístico (Díaz, 2006), en nuestro caso para la etapa de entrenamiento de una red neuronal. En particular permite la detección de fallos en el diseño y toma de datos, el tratamiento y/o la evaluación de datos ausentes, así como la identificación de valores atípicos (Díaz, 2004). Las tareas implícitas en dichas pruebas pueden parecer insignificantes y sin consecuencias a primera vista, sin embargo el impacto que tienen sobre el control de calidad de los resultados puede ser significativo.

Existen diferentes etapas durante el *AED*, las cuales se enumeran a continuación (Figueiras y Gargallo, 2012):

- 1) Preparar los datos para hacerlos accesibles a cualquier técnica estadística.
- 2) Realizar un examen gráfico de la naturaleza de las variables individuales a analizar y un análisis descriptivo numérico que permita cuantificar algunos aspectos gráficos de los datos.
- 3) Realizar un examen gráfico de las relaciones entre las variables analizadas y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre ellas.
- 4) Evaluar, si fuera necesario, algunos supuestos básicos subyacentes a muchas técnicas estadísticas como, por ejemplo, la normalidad, linealidad y homocedasticidad.
- 5) Identificar los posibles casos atípicos (outliers) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.
- 6) Evaluar, si fuera necesario, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

3.2.1 Preparación de los datos

El primer paso en un A.E.D. es hacer accesible los datos a cualquier técnica estadística (Figueiras y Gargallo, 2003), lo cual conlleva a definir el formato del archivo y el paquete estadístico adecuado para procesarlos, así como la codificación del dato según el tipo de variables (tipo carácter, numéricos, categóricos, etc.). En esta etapa los datos pueden ser transformados, editados y ordenados.

En el caso de nuestro problema, después de la organización y control de calidad de nuestros registros de pozo y la clasificación de facies (maestro), se procede generar archivos *ascii*, donde cada uno representa un pozo. El formato de estos archivos se estructura en columnas donde cada una representa curvas de los registros, a excepción de la primera y la última, que representan la profundidad y la clasificación, respectivamente.

En este trabajo de tesis, el análisis estadístico se realizó mediante el entorno de programación para análisis estadístico y gráfico R (Lenguaje R), debido al desarrollo que se tienen en la actualidad de técnicas estadísticas implementadas y su condición de distribución libre. Los datos deben ser estructurados de tal manera que se puedan generar los *dataframes*, que son arreglos de variables de diferentes tipos (escalares, vectores o matrices, ya sean numéricas o cadenas de caracteres).

3.2.2 Análisis estadístico unidimensional y multidimensional

Con los datos organizados en archivos, el segundo paso de un A.E.D. consiste en realizar un análisis estadístico gráfico y numérico de las variables en estudio, con el fin de obtener una idea inicial de la información contenida en el conjunto de datos así como detectar la existencia de posibles errores en los

mismos (Figueiras y Gargallo, 2003). El tipo de análisis a realizar depende de la variable analizada y de su escala de medida (escala nominal, ordinal, de intervalo y de razón).

3.2.2.1 Variables categóricas.

Las variables categóricas son aquellas que no aparecen en forma numérica, sino como categorías o atributos. Dichas categorías pueden tener una escala ordinal o nominal, si es que existe un orden subyacente o no respectivamente (Figueiras y Gargallo, 2003). El tipo de electrofacie (litofacie o clase petrofísica) es un ejemplo de variable categórica con escala nominal.

3.2.2.2 Variables cuantitativas

Las variables cuantitativas son las que pueden expresarse numéricamente y se clasifican en variables cuantitativas discretas, si son resultados de conteos, tomando valores enteros (p.ej. el número de pozos o facies) y cuantitativas continuas, si resultan de una medición, donde pueden tomar valores decimales (p.ej., los valores de los registros geofísicos de pozo).

3.2.2.3 Estadística univariada.

En este trabajo los registros geofísicos de pozo se consideran como variables aleatorias continuas (las facies son ejemplo de variables aleatorias discretas) (Díaz, 2004), por lo que las técnicas gráficas y numéricas utilizadas, intentan bosquejar la función de probabilidad (función de distribución de probabilidad y la función de densidad de probabilidad).

La metodología propuesta considera los siguientes pasos, dentro del análisis estadístico univariado:

- a) Se obtienen y analizan los siguientes estadísticos(estadígrafos): Número de muestras, valor mínimo, primer cuartil, mediana, media, tercer cuartil, valor máximo, rango, rango intercuartil, varianza, desviación estándar, coeficiente de simetría (skewness) y curtosis, los cuales se realizan por variable y pueden ser filtrados por pozo y facie.
- b) De forma gráfica se realizan los histogramas, gráficos de caja y gráficos cuantil-cuantil con el fin de estudiar la forma de la distribución (ver figuras 3.3 y 3.4) y analizar, en particular, la posible existencia de varias modas, que pongan de manifiesto la presencia de diversos grupos homogéneos en la muestra (Figueiras y Gargallo, 2003). Estos gráficos permiten identificar la presencia de valores atípicos y dan una idea sobre la normalidad de la distribución⁷. Los análisis se realizarán para todos los datos, por pozo y por facie, generando interpretaciones específicas para el control de calidad de los registros geofísicos.
- c) Del análisis gráfico anterior, se puede decidir si nuestros datos necesitan ser transformados o no, o si es necesario eliminar los datos atípicos (outliers), buscando mejorar la asimetría de la distribución. Si se opta por transformar los datos o eliminar los valores atípicos (outliers), se debe realizar el análisis gráfico y numérico nuevamente, con el fin de observar el impacto que tienen en la distribución y si es conveniente hacerlo o no.

Como se vio en el capítulo anterior, para mejorar el rendimiento de una red, es recomendable realizar algunas transformaciones, sobre todo para mantener los datos en un rango que facilite el entrenamiento de la red (pre procesamiento), sin olvidar que para el problema de aproximación de funciones, es necesario regresar al dominio de los datos originales (post procesamiento) ante cualquier transformación de los datos. En el caso de problemas de clasificación, no es necesario regresar el dato a su estado original.

⁷ La normalidad de nuestros datos se puede analizar de forma gráfica, mediante qqplots e histogramas, o de forma numérica mediante contrastes de hipótesis.

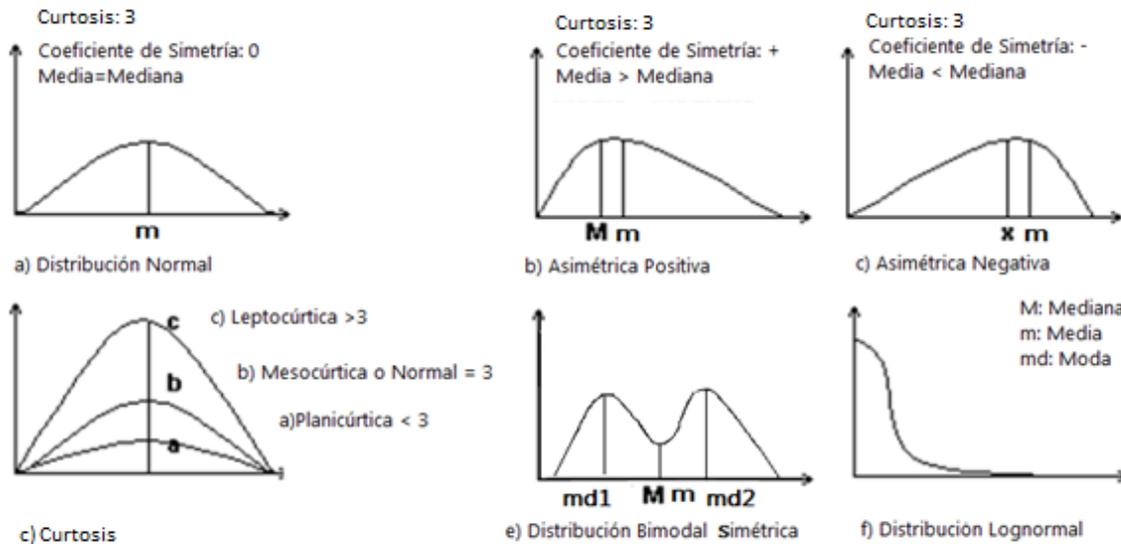


Fig. 3.3
Tipología de las distribuciones de frecuencias agrupadas (modificado de Díaz, 2004)

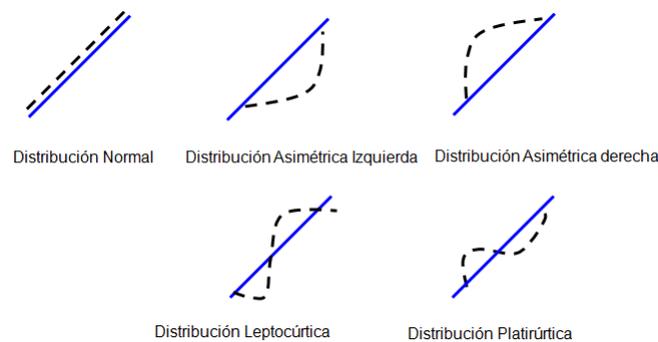


Fig. 3.4
Diagramas de cuantiles correspondientes a distintos tipos de distribuciones (modificado de Salvador Figueras y Gargallo, 2003)

3.2.2.4 Estadística multivariada.

Una vez realizado el estudio unidimensional de cada variable por separado, el siguiente paso consiste en analizar la existencia de posibles relaciones entre ellas. Dicho estudio puede realizarse desde una óptica bidimensional o multidimensional. Éste último podría llevarse a cabo utilizando técnicas multivariantes (análisis de regresión, análisis de agrupamiento, análisis discriminante, análisis factorial y de componentes principales, entre muchas otras). El grado de dependencia entre dos variables aleatorias X y Y, puede ser caracterizado por el diagrama de dispersión o scatterplots alrededor de cualquier línea de regresión (Díaz, 2004) y de los coeficientes de correlación lineales.

En este paso del flujo de trabajo solo se analizan las relaciones entre las variables (registros geofísicos) mediante gráficos de dispersión (en este caso matriz de correlación) y del análisis de la covarianza (Ec. 2.1) y los coeficientes de correlación.

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N x_i y_i - m_x m_y \tag{3.1}$$

Dónde:

σ_{XY} es la covarianza, N el número de datos m_x y m_y , son las medias de las variables aleatoria X y Y, respectivamente.

El coeficiente de correlación es una medida de la asociación que existe entre dos variables, suponiendo generalmente linealidad. Existen diferentes tipos de coeficientes de correlación⁸, cada uno con características específicas, pero el más usado y en el cual se basará este estudio, es el coeficiente de correlación lineal de Pearson (Ec. 3.2).

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad 3.2$$

Donde σ_X y σ_Y son las variables aleatoria X y Y, respectivamente y σ_{XY} su covarianza.

Para modelos lineales el coeficiente de determinación R^2 , mide el grado de la bondad del ajuste, a una línea recta y es igual al coeficiente de correlación al cuadrado.

$$R^2 = r_{XY}^2 \quad 3.3$$

Si $R^2 \approx 1$, el ajuste es bueno (Y se puede calcular de modo bastante aproximado a partir de X y viceversa). Si $R^2 \approx 0$, las variables X y Y no están correlacionadas (linealmente al menos).

3.3 Análisis de Agrupamiento (AA) o Clústeres (AC).

La siguiente etapa de esta metodología para la identificación de electrofacies, es realizar un análisis de agrupamiento (AA), el cual es un conjunto de técnicas encaminadas a encontrar grupos de elementos similares dentro de un conjunto de datos, generando una clasificación de los mismos. Las técnicas del AA construyen subgrupos o clústeres, agrupando los datos de acuerdo a algún criterio específico, tratando de capturar la estructura natural de los mismos, de forma tal que el grado de asociación-similitud entre miembros del mismo clúster sea más fuerte que el grado de asociación-similitud entre miembros de diferentes clústeres, por lo que cada clúster se describe como la clase a la que sus miembros pertenecen.

El análisis de agrupamiento es un método no supervisado, ya que no existen grupos definidos, pues estos se encuentran definidos por los datos mismos. Como se mencionó el objetivo del AC es analizar si los objetos de un grupo son similares o están relacionados entre sí, a partir de criterios establecidos por las variables seleccionadas. Por lo tanto, se hace imprescindible definir una medida de similitud o bien de divergencia para ir clasificando a los individuos en unos u otros grupos.

El esquema puede ser estructurado de la siguiente manera:

- a) Se parte de un conjunto de N individuos caracterizados por n variables.
- b) Se establece un criterio de similaridad.
- c) Se especifica una estructura mediante dendrogramas⁹.
- d) Se escoge un algoritmo de clasificación para determinar la estructura de agrupación de los individuos.

⁸ Coeficientes de correlación lineal de Pearson, de Spearman y la Tau-b de Kendall.

⁹ El dendrograma es la representación gráfica que mejor ayuda a interpretar el resultado de un análisis de agrupamiento.

Es importante resaltar que una buena clasificación de los clústeres, depende de una buena selección de las variables que van a describir a los individuos, del cuidado que se tenga en la selección del criterio de similitud a utilizar y del algoritmo de clasificación seleccionado. Así pues, en las siguientes páginas se dedicará espacio para revisar los criterios más importantes de similitud y los principales métodos de análisis, centrándonos en el algoritmo K-medias.

3.3.1 Criterios de similitud.

Una vez seleccionadas las variables a considerar, cada uno de los individuos sujetos al análisis, serán representados por los valores que tomen estas variables en cada uno de ellos. Para clasificar adecuadamente a los individuos se deberá determinar lo similares o disimilares (divergentes) que son entre sí, en función de lo diferentes que resulten ser sus representaciones en el espacio de las variables (Martínez, 2012).

Existe una enorme cantidad de índices de similitud y divergencia. Todos ellos tienen propiedades y utilidades distintas y habrá que ser consciente de ellas para su correcta aplicación (Martínez, 2012).

Principalmente estos indicadores se basan en la distancia (considerando a los individuos como vectores en el espacio de las variables), o bien, basados en el coeficientes de correlación. En general, el nombre de distancia o disimilitud entre dos individuos \vec{i} y \vec{j} , indicada por $d(\vec{i}, \vec{j})$, mide el grado de semejanza, en relación a un cierto número de características cuantitativa y/o cualitativas. El valor de \vec{i} es siempre un valor no negativo, y cuanto mayor sea este valor mayor será la diferencia entre los individuos \vec{i} y \vec{j} .

Se cumple que:

- a) $d(\vec{i}, \vec{j}) \geq 0$ (no negatividad)
- b) $d(\vec{i}, \vec{j}) = d(\vec{j}, \vec{i})$ (simetría)

Para determinar la homogeneidad entre los individuos se utilizan indicadores de su divergencia (disimilitud) como las distancias; o en su defecto, se pueden utilizar indicadores de la similitud. Existe una gran cantidad de indicadores y no hay una regla general para seleccionarlas, dependerá de la naturaleza de las variables utilizadas y de los individuos estudiados (Martínez, 2012).

3.3.1.1 Distancia euclidiana

La distancia euclidiana es la más conocida y más sencilla de comprender, pues su definición coincide con el concepto más común de distancia.

Su expresión es la siguiente:

$$d(\vec{i}, \vec{j}) = \sqrt{(x_{i1} - y_{j1})^2 + (x_{i2} - y_{j2})^2 + \dots + (x_{in} - y_{jn})^2} \quad 3.4$$

Siendo $\vec{i} = (x_{i1}, x_{i2}, \dots, x_{in})$ y $\vec{j} = (y_{j1}, y_{j2}, \dots, y_{jn})$, puntos definidos en un espacio n dimensional.

3.3.1.2 Otros indicadores de distancia.

Existen otros indicadores de distancia y su uso estará en función de la naturaleza del problema, por mencionar los más importantes se tienen:

1- Distancia euclidiana normalizada:

$$d(\vec{i}, \vec{j}) = \sqrt{\sum_{n=1}^p \frac{(x_{in} - y_{jn})^2}{\sigma_n^2}} \quad 3.5$$

Donde σ_n es la desviación estándar para cada x_{in} .

2- Distancia de Mahalanobis

$$d(\vec{i}, \vec{j}) = \sqrt{(\vec{i} - \vec{j})^T \Sigma^{-1} (\vec{i} - \vec{j})} \quad 3.6$$

Donde Σ^{-1} es la matriz de covarianzas.

3- Distancia de Manhattan:

$$d(\vec{i}, \vec{j}) = \sum_{n=1}^p |x_{in} - y_{jn}| \quad 3.7$$

4- Distancia de Chebyshev:

$$d(\vec{i}, \vec{j}) = \max_{n=1 \dots p} |x_{in} - y_{jn}| \quad 3.8$$

5- Distancia de Minkowski

$$d(\vec{i}, \vec{j}) = \left\{ \sum_{k=1}^p |x_{ik} - y_{jk}|^r \right\}^{\frac{1}{r}} \quad 3.9$$

Se puede definir a la distancia euclídea, como el caso especial de la distancia de Minkowski, donde $r=2$

3.3.2 Criterios basados en medidas de correlación.

Muchos de los indicadores de similitud se basan en coeficientes de correlación o de asociación. El coeficiente de correlación de Pearson se utiliza, en ocasiones, preferentemente con datos cuantitativos y generalmente es socorrido durante la aplicación del algoritmo de mínimos cuadrados. Los coeficientes de correlación por rangos de Kendall y Spearman suelen ser utilizados en variables ordinales.

Si una distancia es grande nos indica que dos puntos son muy diferentes, en cambio los indicadores basados en la correlación, actúan de manera contraria, ya que conforme aumente su valor, aumentará la similitud entre los individuos.

3.3.3 Similitud y distancia entre grupos

La aplicación del análisis clúster requiere, no sólo el cálculo de las distancias o similitudes entre los individuos, sino, también la determinación de las distancias o similitudes entre los grupos y/o entre un grupo y un individuo.

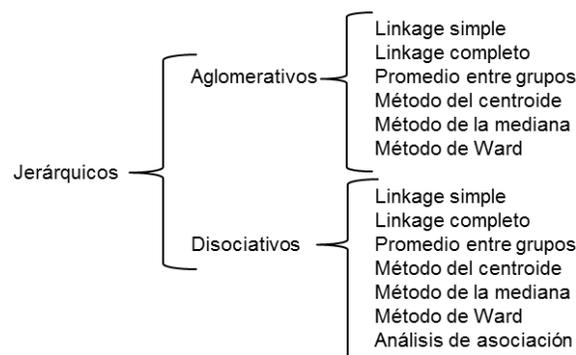
Hay varias alternativas diferentes para definir la distancia entre grupos, las más comunes se enumeran a continuación:

- Distancia mínima (Nearest Neighbour Distance). Se puede utilizar la distancia entre un grupo y un individuo, definida como la menor de las distancias entre los individuos del grupo y el individuo exterior considerado. Este criterio es utilizado para definir la distancia entre grupos.
- Distancia máxima (Furthest Neighbour Distance). También se puede definir la distancia entre un grupo I y un individuo j , como el valor máximo de las distancias entre j y los individuos de I . Al igual que para el inciso anterior este criterio es utilizado para definir la distancia entre grupos.
- Distancia entre centroides. Otro criterio comúnmente utilizado para definir la distancia entre el grupo I y el individuo j , es el de la distancia entre los centroides o centros de gravedad de I y j . Si i es el centro de gravedad de I , tendremos que: $d(I,j) = d(i, j)$ y de la misma manera la distancia entre dos grupos I y J vendrá dada por la distancia entre sus centroides: $d(I,J) = d(i, j)$

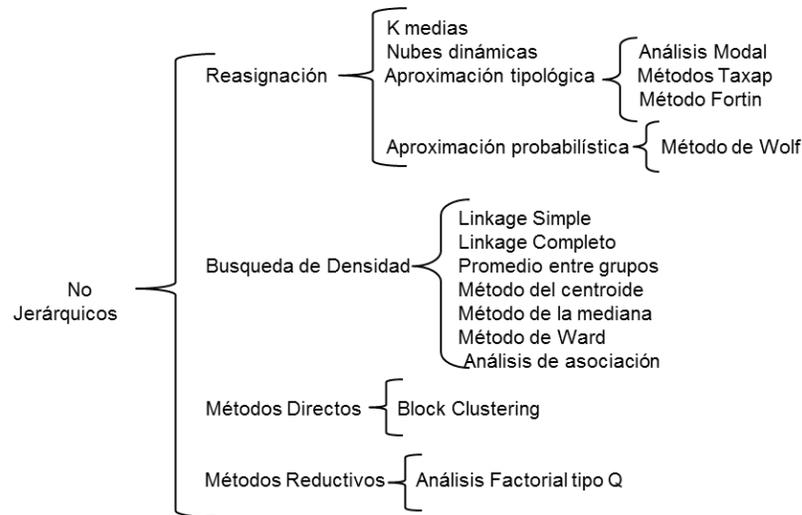
3.3.4 Algoritmos de análisis de grupos

Una vez seleccionada las variables y los criterios de similitud, es necesario escoger el algoritmo que permita formar estos grupos o clústeres. Existen diversos criterios para clasificar los algoritmos utilizados durante el análisis de agrupamiento, comúnmente existe una clasificación muy general de ellos (Cuadros sinópticos 3.1 y 3.2):

- Métodos Jerárquicos.- Agrupan (jerárquicos aglomerativos) clústeres para formar un nuevo clúster o separan (jerárquicos disociativos) alguno ya existente para dar origen a otros grupos de forma que se maximice una medida de similitud o se minimice alguna distancia, siguiendo una jerarquía.
- Métodos No jerárquicos.- Están diseñados para la clasificación de individuos en K grupos. El procedimiento es elegir una partición de los individuos en K grupos e intercambiar los miembros de los grupos para tener una partición mejor.



Cuadro sinóptico 3.1 Clasificación de algunos algoritmos jerárquicos para el análisis de agrupamiento



Cuadro sinóptico 3.2. Clasificación de algunos algoritmos no jerárquicos para el análisis de agrupamiento

3.3.4.1 Método de las k-medias

Es un método que permite asignar cada observación del conjunto de datos, al grupo o clúster que se encuentra más próximo en términos del centroide (media), generalmente usando la distancia euclidiana. La idea principal es definir k centroides situando los puntos en las clases del clúster más cercano, repitiendo este proceso iterativamente (Duda et.al, 2001). El problema del empleo de estos esquemas es que fallan cuando los puntos de un grupo están muy cerca del centroide de otro grupo, también cuando los grupos tienen diferentes tamaños y formas (Karipys et.al, 1999).

Dado un grupo de observaciones $x_p=(x_1, x_2, \dots, x_n)$, donde cada observación es un vector real de dimensión d. El objetivo del método es dividir las n observaciones en k sets de datos S, con $k \leq n$ y

$$S = \{S_1, S_2, S_3, \dots, S_k\} \tag{3.10}$$

Donde k es el número de clúster y S es conjunto de clústeres. Minimizando la distancia euclidiana (Ec. 3.11) se definen dos etapas: la etapa, de asignación a un clúster (Ec. 3.12) y la etapa de actualización (Ec. 3.13)

$$\arg \min_S \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \tag{3.11}$$

$$S_i^{(t)} = \{x_p: \|x_j - m_i^{(t)}\| \leq \|x_j - m_j^{(t)}\| \forall 1 \leq j \leq k\} \tag{3.12}$$

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i} x_j \tag{3.13}$$

Dónde:

$S_i^{(t)}$ = grupo o clúster actual
 $x_j \forall x_p$ = observación j del vector x_p
 μ_i = centroide (media multivariada) del grupo S
 $m_i^{(t)}$ = centroide de la iteración actual
 $m_i^{(t+1)}$ = centroide de la siguiente iteración

Pasos:

1. Se toman al azar k clústeres iniciales
2. Se calculan las distancias o coeficientes de correlación de los centroides, de las variables seleccionadas
3. Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clústeres y se re asignan a los que estén más próximos. Se vuelven a recalcular los centroides de los k clústeres después de las reasignaciones de los elementos.
4. Se repiten los dos pasos anteriores hasta que no se produzca ninguna re asignación, es decir, hasta que los elementos se estabilicen en algún grupo. En la práctica, se observan la mayor parte de reasignaciones en las primeras iteraciones

El método de k-medias necesita como entrada el número de clústeres k. Una división con muchos clústeres complica los resultados porque los hace difíciles de interpretar y analizar. Una división con muy pocos clústeres lleva a una pérdida de información y puede llevar a tomar malas decisiones. Al problema de determinar el número de clústeres se le conoce como "el problema fundamental de la validez del clúster" (Xu, 2005).

Algunos métodos que se han utilizado para encontrar el número adecuado de clústeres son:

- Visualización del conjunto de datos, lo que funciona bien para dos dimensiones pero generalmente nuestros conjuntos de datos son mucho más complicados.
- Mediante diferentes criterios estadísticos: criterio de Información de Akaike (AIC), criterio de Inferencia Bayesiana u Optimización de otras funciones de criterio.
- Otros métodos heurísticos basados en una variedad de técnicas y teorías.

Se utiliza este método para el problema específico de estimar las litofacies, por su rápida aplicación y popularidad en la clasificación de litofacies (Wolf y Pelissier-Combescure, 1982; Faria., et.al, 2005 y Popielski, et. al., 2012)

3.3.5 Análisis de agrupamiento en la clasificación de electrofacies.

El objetivo del análisis de agrupamiento en nuestro problema específico de clasificación de electrofacies, es identificar la naturaleza de los datos y la relación que guarda el conjunto de registros en un conjunto de entrenamiento. Las diferentes técnicas del AA resuelven problemas de clasificación al igual que las redes neuronales, sin embargo su utilidad en este estudio se justifica por las siguientes razones:

- a) Tener una idea inicial del número de clases, desde un punto de vista cuantitativo y función únicamente de los datos de registros geofísicos de pozo.

- b) Se inicia probando con el número de clases propuesta en la interpretación manual, para posteriormente variar aleatoriamente el número de clústeres k (e pueden utilizar criterios como el de información de Akaike)
- c) Para redes supervisadas, el AA, permite tener una idea de la consistencia numérica de los datos de los patrones de entrenamiento de salida (maestro o guía), con la clasificación propuesta por el experto petrofísico.
- d) En el caso de una red no supervisada, permite tener una idea del número de clases que se tendrían que definir al momento de entrenar la red.
- e) Tener un punto de comparación de los resultados.

3.5 Análisis de selección de variables.

En el análisis de selección de variables, se definen cuales variables serán utilizadas para maximizar el entrenamiento de la red neuronal, mediante el análisis de diferentes pruebas estadísticas, como por ejemplo: análisis exploratorio de datos, análisis de agrupamiento, análisis discriminante, componentes principales, entre muchas otras; así como la combinación de varias de estas técnicas.

Como se mencionó en el capítulo anterior, un mayor número de variables de entrada, implica un mayor número de unidades de procesamiento (neuronas), lo cual, en lugar de ser beneficioso implica un mayor cantidad de parámetros libres, que conducirá a un entrenamiento más deficiente y la posibilidad de conducir a problemas en la capacidad que tiene la red de generalizar; por lo que es altamente recomendable utilizar solo la información que sea necesariamente útil y que contribuya más para la solución del problema en cuestión. En este trabajo se propone el uso de la técnica de componentes principales (CP), debido a que permite reducir el número de variables, mediante un preproceso, utilizando solo la información que tiene mayor correlación entre sí; así como también identificar y reducir el efecto que tienen los datos atípicos multivariantes.

3.4.1 Análisis de componentes principales (ACP).

Uno de los problemas centrales en el análisis de datos multivariantes es la reducción de la dimensionalidad. En el análisis de componentes principales se realiza una transformación de variables, por medio de una rotación ortogonal (y/o una translación), ya que dado un conjunto de n observaciones de p variables puede representarse en el espacio p dimensional como la proyección sobre un hiperplano (sub espacio vectorial) de dimensión $m < p$, que contenga la mayor cantidad puntos del conjunto de datos. Lo cual es equivalente a encontrar un nuevo sistema de coordenadas que represente lo mejor posible los datos sin causar distorsiones, lo anterior equivale, si es posible, a representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales (Peña, 2005).

La técnica de componentes principales se debe a Hotelling en 1933, aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson en 1901. La utilidad de esta técnica es doble:

1. Permite representar óptimamente en un espacio de dimensión pequeña observaciones de un espacio general p -dimensional (Fig. 3.5). En este sentido, las componentes principales son el primer paso para identificar las posibles variables observadas latentes o no, que generan los datos.
2. Permite transformar las variables originales, en general correlacionadas, en nuevas variables incorrelacionadas, facilitando la interpretación de los datos.

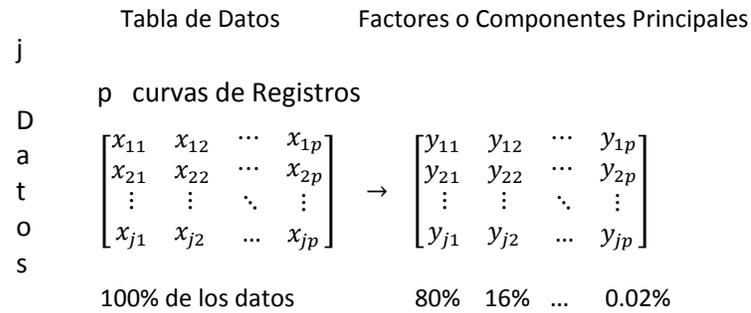


Fig. 3.5. Transformación de las variables originales en factores o componentes principales (modificado de Rodríguez, 2009)

Un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante y por tanto, pocos factores explicarán gran parte de la variabilidad total. Es importante resaltar el hecho de que el concepto de mayor información se relaciona con el de mayor variabilidad o varianza.

Si las variables originales están incorrelacionadas de partida, entonces no tiene sentido realizar un análisis de componentes principales. El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes.

La elección de las componentes principales o también llamados factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente (Terrádez, 2012). En el caso de los registros de pozos, se desea convertir p curvas de registros con j muestras cada uno (Fig. 3.5), en variables y_{jp} , que representen un porcentaje de los datos originales (en el ejemplo de la Fig. 3.1 y_{j1} representa el 80%, y_{j2} el 16%, etc.)

Si se calcula la matriz de covarianzas Σ o la matriz de correlación (matriz de coeficientes de correlación de Pearson), entre las p variables, existe una matriz ortogonal A (matriz de rotación), de tal manera que:

$$\Sigma = ADA^{-1} \tag{3.14}$$

Donde D es la matriz diagonal formada por los valores propios o factores de Σ , denotados como $\lambda_1, \lambda_2, \dots, \lambda_p$. En el ACP estos factores, se ordenan de tal manera que $\lambda_1 < \lambda_2 < \dots < \lambda_p$, lo cual es simplemente el reordenamiento de las columnas de la matriz A, de forma que la primera columna sea el vector propio normalizado o componente asociado a λ_1 , la segunda el vector propio asociado a 2 y así sucesivamente, siendo perpendiculares entre sí y de longitud igual a la unidad.

Si

$$A = a_{ip} \tag{3.15}$$

Tenemos que $a_{1p} = (a_{11}, a_{12}, \dots, a_{1p})'$ apunta a la dirección de máxima variabilidad del conjunto de datos y a_{2p} , apunta a la siguiente dirección y así sucesivamente.

Cada elemento de ésta matriz, representará los coeficientes factoriales de las variables. La matriz tendrá tantas columnas como componentes principales y tantas filas como variables (Terrádez, 2012).

Si se considera una serie de variables x_i (x_1, x_2, \dots, x_p) y se trata de calcular, a partir de ellas, un nuevo conjunto de variables y_1, y_2, \dots, y_p , incorrelacionadas entre sí, cuyas varianzas vayan decreciendo progresivamente, se tendrá que cada y_j (donde $j = 1, \dots, p$) es una combinación lineal de las x_1, x_2, \dots, x_p originales, es decir:

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \dots + a_{jp}x_p = a'_j x \quad 3.16$$

Siendo $a_{jp} = (a_{j1}, a_{j2}, \dots, a_{jp})$ un vector de constantes, los cuales son los vectores propios asociados a x_i (Ec. 3.17)

$$x_i = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \quad 3.17$$

Generalmente la primer componente principal, y_1 , no es suficiente para condensar la información contenida en x , por lo que se construye una segunda componente principal y_2 , luego una tercera y_3 y así sucesivamente. En general en la etapa k , se construye la componente principal p -ésima.

Los factores a_{kj} constituyen un sistema de pesos para las variables, los cuales indican cuanto aporta cada variable a la construcción de cada componente. La idea de la técnica es maximizar la varianza (mayor información) en las componentes, de tal manera que cumplan ciertas restricciones como incorrelación entre componentes (conforme se calculen más componentes principales, la varianza irá disminuyendo progresivamente), ortogonalidad y dimensionalidad. Existen diversos trabajos relacionados con la técnica de las componentes principales, para más detalles y sobre la técnica y el cálculo de los factores puede revisarse el trabajo Jolliffe (2002).

Existen reglas para saber el número más adecuado de componentes principales a conservar, por ejemplo (Bernal, 2009):

- a) El criterio de Kaiser que indica que hay que conservar los componentes principales cuyos valores propios son mayores que la unidad.
- b) Seleccionar las componentes con valores propios (varianzas) mayores que el valor propio (varianza) promedio.
- c) Usar el "test scree" o gráfico de sedimentación, en el que se representa el número de orden del valor propio (j) de y_j y su magnitud. Se descartan las componentes que ya no aportan significativamente a la varianza (Fig. 3.6).
- d) El criterio más utilizado es el de observar el porcentaje de varianza total explicada por cada componente o factor y cuando éste llega a un porcentaje acumulado considerado alto, normalmente cerca del ochenta por ciento, significa que el número de factores es suficiente (Fig. 3.6)

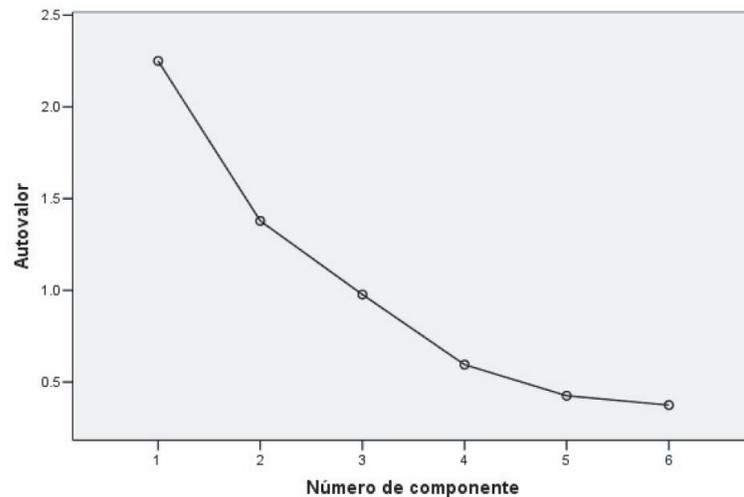


Fig. 3.6. Gráfico de sedimentación tomado de León et. al (2008)

3.4.1.1 Análisis de componentes principales en la estimación de electrofacies.

El análisis de componentes principales ha ayudado en los estudios de clasificación de litofacies (Ma, 2011; Enikanselu y Ojo, 2011; Crain, 2012), permitiendo a técnicas de clasificación de datos, como las redes neuronales, facilitar su entrenamiento, ya que un menor número de variables, implica un menor número de unidades de procesamiento (neuronas), optimizando los resultados.

En la metodología de este trabajo, se propone el uso de las componentes principales, ya que permitirá conocer las variables (registros de pozos) que responden mejor a la clasificación propuesta de electrofacies y así plantear un esquema de entrenamiento. Esta técnica es una herramienta de pre-procesamiento de datos muy útil durante el entrenamiento de los sistemas supervisados y no supervisados, ya que facilita los cálculos y generalización de la red.

Se propone el uso de herramientas como Matlab o R, para realizar este análisis, sin embargo plataformas como petrel (Schlumberger, 2009), Interactive Petrophysics (Senergy, 2005), entre otras, tienen implementado esta herramienta como parte del análisis de selección de variables antes de realizar estimaciones con redes neuronales.

En este trabajo metodológico, el proceso de análisis de componentes principales se propone de la siguiente manera:

1- Se recomienda Normalizar los datos con respecto a la varianza (función mapstd de pre-procesamiento en Matlab), con el fin de trabajar adimensionalmente y con ello usar indistintamente la matriz de correlaciones o de covarianza. La normalización de los datos se realiza de la siguiente manera:

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{\sigma_i} \text{ con } i=1,2,\dots,p \quad 3.18$$

Donde x_i es la variable, \tilde{x}_i es la variable normalizada, \bar{x}_i es el promedio de esa variable y σ_i es la desviación estándar de esa variable.

2- Análisis de la matriz de correlaciones. Un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante y por lo tanto, pocos factores explicarán gran parte de la variabilidad total.

3- Selección de los componentes. La elección se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo recogerá la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Se deberán elegir aquéllos que recojan el porcentaje de variabilidad que se considere suficiente (generalmente superior al 90%). Se calcularán dichos porcentajes y se analizarán los gráficos de sedimentación, para decidir el número adecuado de componentes principales

4- Para seleccionar las variables, de estudio, se deben interpretar los resultados, en función del problema en estudio (correlación entre registros de pozo) y en función de este análisis; con lo cual se decidirán cuáles son las componentes que serán utilizadas para entrenar la red neuronal.

3.5 Diseño y entrenamiento de la red neuronal.

Para esta metodología se propone el uso de registros geofísicos de pozo, que respondan mejor a una clasificación determinada y que son elegidos de los análisis estadísticos previos. Para el diseño, entrenamiento de redes neuronales y clasificación de electrofacies, se propone un programa diseñado en lenguaje M de Matlab (perceptrón multicapa, mapas auto-organizados y committees machines), las gráficas de los resultados son realizadas en Petrel.

Durante la etapa del diseño y el entrenamiento de la red, se determinan una serie de parámetros, que permitirán crear la red neuronal adecuada, para el problema en cuestión y permitirá obtener mejores resultados al momento de presentar los datos a la red. Es importante conocer la estructura que tendrán cada uno de los modelos, la forma en que serán presentados los datos y los algoritmos de entrenamiento para cada modelo neuronal.

Para diseñar de manera más adecuada la estructura de una red neuronal, optimizar sus parámetros y otros factores que afectan su desempeño, se proponen pruebas de ensayo y error, mediante el uso de registros sintéticos, ya que se pueden controlar para casos específicos previamente definidos.

Durante la etapa de diseño y entrenamiento de las redes es importante analizar y definir:

1. Topología
2. Los pesos y bias iniciales
3. Funciones de activación.
4. Los datos de entrada (Entrenamiento y prueba) y la forma de ser presentados a la red.
5. Algoritmos de entrenamiento y sus parámetros.

3.5.1. Topología.

Como se mencionó en el capítulo 2, las arquitecturas de redes neuronales propuestas en este trabajo son: Las redes perceptrón multicapa (*MPL*), los mapas auto-organizados (redes de Kohonen) y los sistemas de combinación modulares y de conjunto (committee machines), por lo que es importante determinar la estructura que tendrán cada una de ellas, la forma en que serán presentados los datos y los parámetros específicos de los algoritmos de entrenamiento para cada modelo neuronal. Los parámetros a determinar en esta etapa son: El número de capas (*MPL* y committee machines), el número de

neuronas por capa (SOM^{10} , MPL y CM) y el número de redes que constituyen un sistema modular y de conjunto¹¹ (CM)

No existe una regla clara para definir las características de la estructura de la red, ya que depende de los datos de entrada y los objetivos a seguir, sin embargo, como se mencionó en el capítulo 2, la idea central es optimizar la generalización del modelo, por lo que considerar un gran número de capas y de neuronas, provocaría que una red neuronal “memorice”.

3.5.2 Pesos y bias iniciales.

Los métodos de entrenamiento de las redes neuronales basados en el descenso del gradiente, son técnicas de búsqueda local que tienen la posibilidad de quedar atrapadas en zonas de la superficie de error, en donde el gradiente es nulo, condición que asegura que se ha encontrado un mínimo, pero no asegura que sea un mínimo global, por lo que la determinación de los pesos iniciales será muy importante en la definición de la calidad del mínimo encontrado y el número de iteraciones requeridas en la búsqueda (Sánchez, 2009).

En la práctica no existe criterio definido para asignar valores a los pesos y los bias iniciales, generalmente se suele iniciar aleatoriamente, por lo que cada vez que la red sea reentrenada, estos serán reinicializados automáticamente con un valor distinto, por lo que se obtendrán resultados diferentes en cada corrida. Usualmente algunas aplicaciones como Matlab, permiten reinicializar los pesos aleatoriamente, bajo la misma secuencia de números aleatorios o definirlos.

La asignación de valores aleatorios a los pesos de una red es un método ampliamente utilizado en la inicialización, sin embargo, diversos investigadores como Nguyen y Widrow (1990) han demostrado su inconveniencia. Iniciar el proceso de entrenamiento con valores aleatorios para los pesos y el bias, puede conducir a una situación en la que numerosas neuronas contribuyan muy poco al desempeño de la red y su actualización durante el proceso de entrenamiento sea inútil y lenta de forma innecesaria, con la inseguridad de que las neuronas restantes sean suficientes para representar adecuadamente el sistema modelado (Sánchez, 2009).

La elección de los pesos y los bias influyen en lograr que la red llegue a encontrar el óptimo global o no, así como en la rapidez de convergencia. Debido a las características de las funciones de activación, es necesario evitar que los pesos, las activaciones de las neuronas o sus derivadas, se hagan próximos a cero, por lo cual, los pesos y bias iniciales no deben de ser muy grandes, ni muy bajas (Sánchez, 2009).

Otra forma de inicializar los pesos y los bias es mediante el algoritmo de Nguyen y Widrow, el cual asume que los pesos se ajustan, dividiendo la región a predecir en pequeñas zonas. Por lo que para acelerar la convergencia del proceso, se asigna desde el principio a cada uno de los pesos, valores en su propio intervalo de inicio del entrenamiento. Estos intervalos individuales se traslapan ligeramente y durante el proceso de entrenamiento se van reajustando en su rango de acción. El algoritmo de Nguyen y Widrow tiene la siguiente secuencia:

$$\beta = 0.7 p^{\frac{1}{n}} \quad 3.19$$

Donde p es el número de nodos ocultos y n es el número de entradas a la red. Para cada neurona oculta j se determina el peso preliminar w_{ij} ($prel$) con todas las neuronas de entrada i , aleatoriamente en el rango $[-0.5, 0.5]$ (Sánchez, 2009).

¹⁰ Los mapas auto-organizados o redes de Kohonen, constan únicamente de una sola capa.

¹¹ La cantidad de redes neuronales será determinado por los datos y la forma en que se decida presentar los mismos.

Se determina

$$\tau = \sqrt{\sum_{i=1}^{i=n} w_{ij} (prel)} \quad 3.20$$

Inicializando cada peso mediante:

$$w_{ij} = \frac{\beta w_{ij} (prel)}{\tau} \quad 3.21$$

La inicialización del bias, se realiza con un valor determinado aleatoriamente en el rango $[-\beta, \beta]$

3.5.3 Funciones de activación

Existen diferentes funciones de activación, las más utilizadas se mostraron en el capítulo anterior, sin embargo, no se limitan a ellas únicamente, pueden utilizarse algunas definidas por el analista, en función de sus necesidades (funciones seno, coseno, etc.).

El criterio más práctico para definir la función de activación para las neuronas de cada capa, es principalmente asociado al problema en cuestión. Las funciones de activación tienen que estar acordes con el tipo de número que van a recibir en la capa de entrada. Si reciben algo en $[0,1]$ se pueden usar las funciones logsig o pureline, si es algo entre $[-1,1]$ se puede usar la función tansig. Lo mismo sucede con la capa de salida: si se quiere devolver algo en $[-1,1]$ se utiliza la función tansig y si es entre $[0,1]$, utilizaremos la función logsig. La función Purelin recibe y devuelve un intervalo más abierto ya que la entrada a la neurona será prácticamente su salida, por lo que el rango de valores es más amplio. En la industria las funciones más utilizadas son hardlims, purelin, logsig y tansig (ver capítulo 2).

3.5.4 Datos de entrada

Uno de los pasos primordiales al aplicar redes neuronales de manera general, es la manera en que los datos son presentados a la red durante el entrenamiento, ya que esto definirá su rendimiento y por ello influirá directamente en los resultados. Como se mencionó en el capítulo 2, el objetivo de del entrenamiento es desarrollar la capacidad de la red neuronal de aprender y generalizar, evitando que la red memorice y sea solamente capaz de reproducir los datos con que fue entrenada.

Una de las formas más habituales de presentar los datos a una red supervisada (red *MPL*), es mediante los “sets de entrenamiento y prueba”, los cuales se crean muestreando un porcentaje determinado de los datos de entrada (puede o no ser aleatoriamente), tanto para el entrenamiento de la red (conjunto de entrenamiento) y utilizar los datos restantes (conjunto de prueba) para probar la red.

La idea global es que con el conjunto de entrenamiento se calculen iterativamente las salidas de la red neuronal, se comparen con el patrón de entrenamiento de salida (maestro o guía) y se establezca una función de error, la cual será minimizada mediante el ajuste de los pesos sinápticos a través de los algoritmos de entrenamiento. Con el conjunto de entrenamiento se pretende calcular las salidas de la red, con datos que no han sido presentados anteriormente a la red, durante el entrenamiento y con ello monitorear el rendimiento de la red durante las diferentes iteraciones y epochs. Este proceso se realiza un determinado número de epochs (e iteraciones), hasta que el error decaiga o hasta que un determinado número de epochs, definido por el analista, se haya alcanzado.

En el caso de las redes no supervisadas, no es común la división en sets de entrenamiento y prueba, debido a la capacidad de este modelo para generalizar, sin embargo, algunas aplicaciones como

SOMPAK (Kohonen et.al., 1995) utilizan sets de prueba para monitorear el rendimiento de la red neuronal.

Para los modelos de conjunto y modulares (Committee Machines), Alpana (2009) propone establecer una topología fija (número de capas y neuronas) y entrenar cada una de las redes perceptrón multicapa que forman la máquina comité (Figs. 3.7, 3.8, 3.9 y 3.10), con sets de nuevos generados a partir del conjunto de datos original.

En Alpana (2009) se aprovecha que los registros de pozo pueden ser tratados como series de tiempo (donde el tiempo t es la profundidad), para generar los distintos sets de datos con desplazamientos en profundidad. En este trabajo de tesis, se propone entrenar cada una de las redes perceptrón multicapa que conforman las máquinas comité (CM), únicamente con los sets de entrenamiento, validación y prueba, muestreados de manera aleatoria.

3.5.5 Algoritmos de entrenamiento

Como se observó en el capítulo 2, cada uno de los modelos propuestos (*MPL* y *SOM*), tienen reglas teóricas definidas para la actualización de los pesos y cada una de ellas pueden ser resueltas por una gran cantidad de algoritmos numéricos definidos. La mayoría de estos algoritmos, utilizan el principio del algoritmo de retropropagación (backpropagation), para optimizar la función objetivo o función de error.

Cada uno de los algoritmos numéricos tiene sus propias características y sus propios parámetros. La descripción y entendimiento de los mismos, queda fuera del alcance de este trabajo. La elección del algoritmo y la definición de sus parámetros, no es una tarea sencilla y con reglas específicas, generalmente dependiendo del entendimiento del problema se pueden establecer relaciones empíricas, no obstante su aplicación, se limita a ensayos de prueba y error.

Es importante señalar que aplicaciones como Matlab, permiten escoger entre una amplia gama de funciones de entrenamiento y de aprendizaje (Hagan, et. al., 2009). El uso de las mismas estará en función de los resultados obtenidos con las pruebas de los datos sintéticos.

3.5.6 Entrenamiento

En la etapa de entrenamiento se describe el proceso global de la interacción de la estructura del modelo neuronal utilizado (topología), la forma en que los datos se presentan a la red, la inicialización de los pesos y los bias, así como los algoritmos de entrenamiento.

En este trabajo se pretende realizar una clasificación de facies de forma binaria (ceros y unos), por ejemplo, si a una profundidad determinada la facie A existe, la salida será 1, si no existe será 0.

En la etapa de entrenamiento, se realizarán, todos los ajustes necesarios para optimizar el proceso de clasificación.

3.5.6.1 Entrenamiento de la red perceptrón multicapas

Para la red perceptrón multicapa (*MPL*), se realiza el entrenamiento probando diferentes topologías y muestreando diferentes conjuntos de datos de entrenamiento, validación y prueba (variables analizadas estadísticamente y pre-procesadas). Durante el diseño de la red neuronal, se definen los diferentes parámetros que ofrece MATLAB, tanto para inicializar los pesos y los bias, así como para actualizarlos (Algoritmo de entrenamiento).

3.5.6.2 Entrenamiento de los mapas auto-organizados.

En el caso de las redes de Kohonen, se realiza el mismo esquema de las redes perceptrón multicapa, enfatizando los análisis estadísticos previos, principalmente el número de clases obtenido por el análisis de agrupamiento, ya que es un parámetro de entrada que restringe la construcción de los clústeres durante la clasificación. Lo anterior implica un conocimiento previo del sistema de depósito y el número de electrofacies que se esperarían obtener. Los mapas auto-organizados, permiten que los datos se agrupen naturalmente en clases, sin ninguna restricción, dependiendo únicamente del espacio de características (espacio de entrada). En Matlab existen algunas librerías (*SOM Toolbox* de Matlab), que permiten restringir el número de clústeres y permiten definir los sets de entrenamiento, validación y prueba. “En este trabajo de tesis, se propone analizar la clasificación generada por los datos auto-organizados, sin ninguna restricción”

Otros parámetros a definir durante el entrenamiento de mapas auto-organizados, son la dimensión de la malla (Matlab permite mapear en 1 y 2D), así como el número de neuronas y su distribución topológica.

3.5.6.3 Entrenamiento de las máquinas comité (Committee Machines)

En Alpana (2009), se menciona que el objetivo del enfoque de conjunto (ensemble) es intentar mejorar el rendimiento. La estrategia de entrenamiento en este trabajo de tesis, implica que cada red MLP que conforman la red de conjunto (Fig. 3.7), pueda ser entrenada para generalizar de formas diferentes (lo cual implica modificar sus parámetros de entrenamiento, como topologías, épocas, factor de aprendizaje y parámetros propios de los algoritmos de entrenamiento)

La metodología de este trabajo se propone bajo el siguiente esquema:

1. El espacio de entrada será una matriz cuya dimensión será de m muestras \times n variables o registros (Fig. 3.7)
2. La dimensión del espacio de salida de la red perceptrón multicapa, será binario (Fig. 3.7).
3. Se propone descomponer la tarea de clasificación de electrofacies, en subtarefas realizadas por módulos de redes neuronales, donde cada módulo resolverá una electrofacie en particular.
4. Cada módulo (Fig. 3.9) tiene por entradas los registros geofísicos de pozo y las salidas de las red de conjunto (ensemble neural network) (Fig. 3.8).
5. Una red neuronal de conjunto (Fig. 3.8), es un conjunto de redes perceptrón multicapa, cuyas salidas alimentan a otra red perceptrón multicapa (Fig. 3.7), que proporcionara la salida global del conjunto de redes neuronales, para alimentar posteriormente al módulo (Fig. 3.9)
6. El Número de redes perceptrón multicapa simples k (Fig. 3.7), depende de los datos de entrada y/o de las pruebas realizadas en los datos sintéticos.
7. El conjunto de todos los módulos formarán una Maquina Comité (Committee Machine) o red neuronal modular (modular neural network), la cual es la combinación de las salidas de i módulos (descritos en el punto 4) mediante un proceso de combinación (combinador), para obtener la clasificación final de las electrofacies (Fig. 3.9).
8. El combinador permite obtener la salida final de las salidas específicas de cada uno de los módulos. Existen algunas formas de combinar las salidas de los diferentes expertos, mediante algoritmos o técnicas como: algoritmos genéticos, técnicas difusas, clasificadores dinámicos,

combinaciones lineales, redes neuronales, etc (Yu et. al., 2003). En este trabajo se utilizarán las redes de Kohonen como combinadores.

9. Durante el entrenamiento de las redes de conjunto, se propone probar diferentes modelos de redes perceptrón multicapa, con parámetros y topologías distintas (definidos mediante pruebas), alimentadas con datos de entrada muestreados aleatoriamente.



Fig. 3.7. Esquema red neuronal perceptrón multicapa

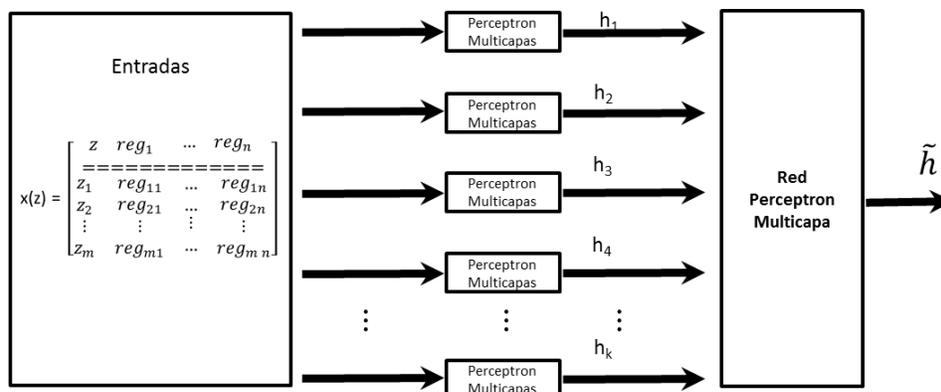


Fig. 3.8. Esquema red neuronal de conjunto.

3.6 Análisis del error y validación cruzada (durante la etapa de entrenamiento)

Como se ha venido mencionando hasta el momento, el objetivo de los modelos neuronales, es desarrollar su capacidad de generalizar, como se mencionó en el capítulo anterior, esto se logra dividiendo el conjunto de datos de entrada en sets de datos para entrenamiento y prueba. Durante el entrenamiento, es necesario analizar el rendimiento que la red pueda tener. La manera de hacerlo es comprobar su capacidad de predicción durante cada una de las etapas del entrenamiento.

La manera de medir la capacidad del predictor es mediante metodologías de validación, por ejemplo:

- a) Holdout: Se divide el conjunto de casos en dos grupos: conjunto de entrenamiento (2/3) y conjunto de prueba (1/3). El conjunto de entrenamiento se usa para generar el clasificador y el de test para evaluarlo.
- b) Validación cruzada (cross-validation): Se divide el conjunto de datos en K subconjuntos del mismo tamaño. Se utilizan K-1 subconjuntos como datos de entrenamiento y 1 subconjunto como datos de test. Se repite para los K subconjuntos y se calcula la media de la evaluación. Suele utilizarse K=10.
- c) Dejar uno fuera (leave one out): Validación cruzada con K igual al número de datos (K más grandes). Este método es ideal con funciones de error continuas, como el MSE (mean square error), pero puede encontrar dificultades al utilizarse para funciones discontinuas (además de las evidentes en cuanto a tiempo de computación).

d) Jackknife: Esta técnica es comúnmente confundida con el método de validación cruzada, sin embargo se usa para calcular el error sistemático (bías) de una variable estadística de interés en cada subconjunto de datos

e) Bootstrapping: El conjunto de entrenamiento se escoge como una muestra aleatoria con reemplazamiento.

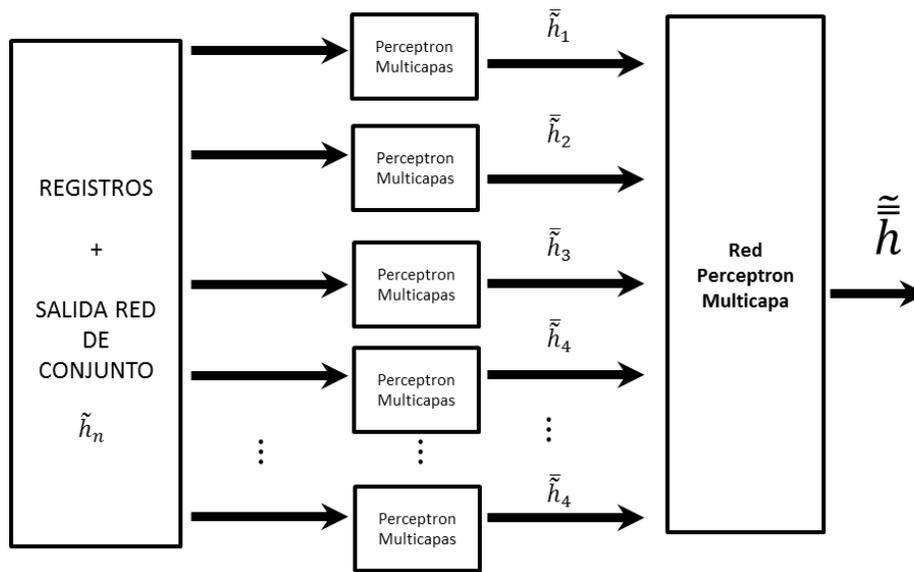


Fig. 3.9 Esquema de un módulo

En Matlab generalmente se utilizan los métodos de Holdout, leave one out y validación cruzada, donde el tamaño y el tipo de muestreo (aleatorio o no) pueden ser definidos. Como se ha venido manejando desde el capítulo 2, en Matlab se definen tres grupos de conjunto de datos: conjunto de entrenamiento, conjunto de prueba y un nuevo conjunto de datos comúnmente llamado conjunto de validación (del cual ya se ha venido hablando).

El objetivo de dividir los datos de entrada en diferentes sets, persigue evaluar el error de la red tras cada epoch (o tras cada cierto número de epochs) y determinar el momento en que éste empieza a aumentar. El conjunto de entrenamiento será utilizado para modificar los parámetros de la red. El conjunto de validación es independiente del entrenamiento (ajuste de pesos), pero prueba el modelo durante el entrenamiento. El conjunto de prueba no contribuye al proceso de entrenamiento en nada y sirve para determinar si la red puede estimar correctamente las salidas, ante datos que no fueron presentados antes a la red.

El proceso de aprendizaje de la red previamente descrito, ocurre epoch a epoch hasta que los pesos y los bias (factor de polarización) se estabilizan y que el criterio de error (típicamente el error cuadrático medio (*MSE*)) converge a algún valor mínimo. Además, es conveniente que las muestras se presenten de forma aleatoria para convertir en estocástica la búsqueda en el espacio de los pesos y así minimizar la posibilidad de convergencia a un mínimo local.

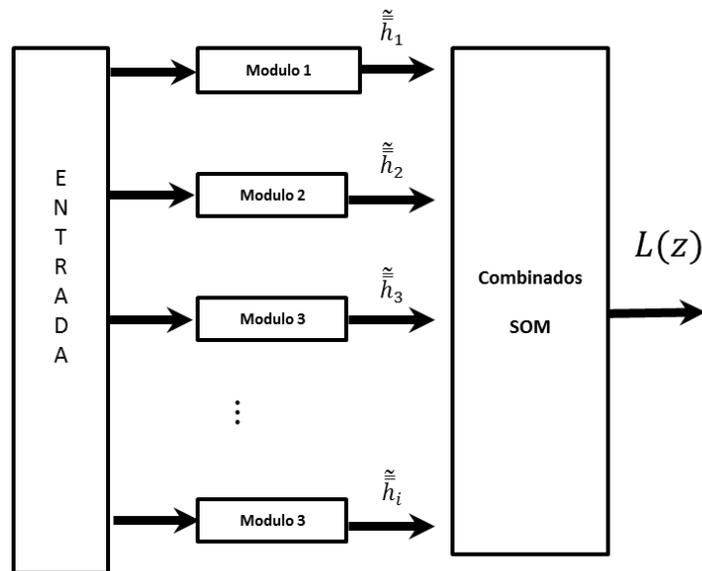


Fig. 3.10 Esquema de la Máquina Comité (Committee Machine) o red neuronal modular (conjunto de módulos)

Para monitorear la red supervisada (*MLP* y *CM*) y evaluar su desempeño, se recomienda realizar el siguiente análisis para el error.

1.- Para cada epoch se calculan los errores cuadráticos medios, que el modelo produce al estimar la salida de la red para los sets de entrenamiento, validación y prueba (Matlab lo realiza de forma automática).

2.- Se procede a graficar el número de epochs contra los errores cuadráticos medios de las estimaciones de los sets de entrenamiento, validación y prueba (grafico del performance - Fig. 3.11a), para monitorear el rendimiento durante el entrenamiento. Bajo este criterio, se puede observar en el gráfico, que el número de epoch óptimo es cuando se alcanza el error cuadrático medio mínimo en el conjunto de validación.

3.- Se generan los gráficos de confusión (Fig. 3.11b), los cuales, son herramientas de visualización, donde las columnas de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa los elementos reales que pertenecen a cada clase. En estos gráficos se muestran los porcentajes totales, por clase y global.

4- Para problemas de clasificación, no se realiza el análisis exploratorio del error, debido a que para este caso, las salidas se encuentran entre uno y cero.

Para los mapas auto-organizados, Matlab no considera dividir los datos en sets de entrenamiento, validación y prueba; sin embargo, puede realizarse esta división y generar un proceso de validación cruzada, si se conoce la salida esperada. En este trabajo de tesis se propone únicamente monitorear el error mediante el uso de matrices de confusión, cuando exista una clasificación previa para comparar.

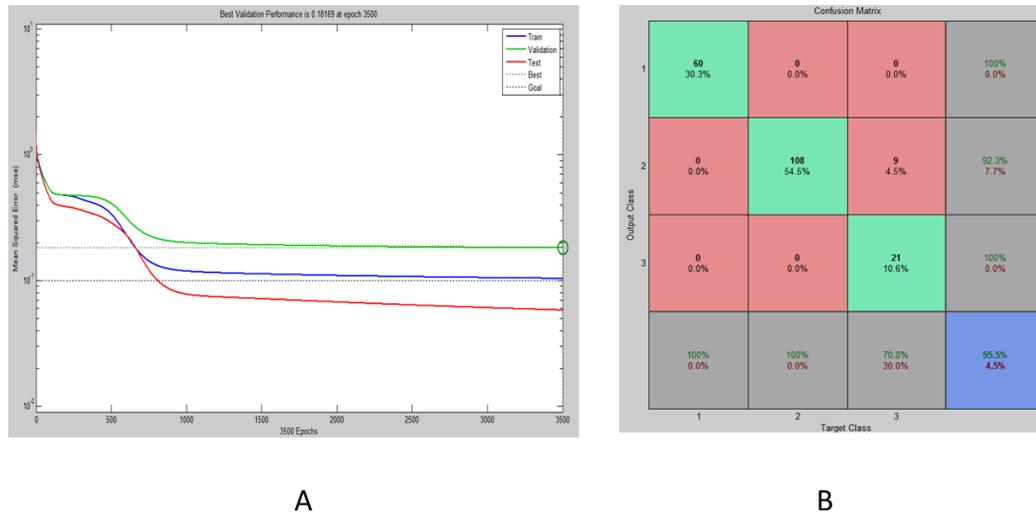


Fig. 3.11. Ejemplos de: a. Graficas de performance, b. histograma del error absoluto y c. matriz de confusión

Para el proceso de entrenamiento, Matlab cuenta con una serie de herramientas gráficas, que conviene visualizar, para verificar el rendimiento del entrenamiento:

1- Mapa de Éxito (Hit Plot).- El mapa de éxito (Fig. 3.12), muestra la topología de la red y el número de datos de entrada en cada clase (el número de datos reconocidos por cada neurona). Áreas con números mayores de éxito indican clases que representan regiones altamente pobladas con características similares, del espacio de entrada, mientras que números bajos de éxitos muestran regiones espaciadamente pobladas (Hagan, et. al., 2009).

Las neuronas con ceros, pueden interpretarse como los límites bien definidos entre clústeres.

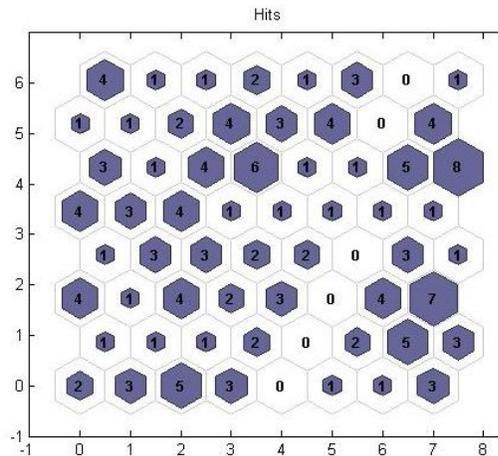


Fig. 3.12 Ejemplos Gráfico de éxitos, topología de 8x8 Manual de Matlab (2009)

2- Mapa de distancia de pesos entre vecinos (Neighbor Weight Distances).- También llamada matriz-U, muestra la distancia (distancia euclidiana, utilizando los pesos) a la que se encuentra cada clase de neurona de sus vecinos, así como las conexiones topológicas. La escala de colores indica que zonas más brillantes, son áreas altamente conectadas y relacionadas en el espacio de entrada. Mientras que conexiones con tonos más oscuros indican regiones con pocos o ningún dato reconocidos por esas neuronas (Manual de Matlab, 2009). En la Fig. 3.13, se puede observar que zonas con colores de rojo oscuro a negro, marcadas por los polígonos verdes limitan las áreas con características distintas

(diferentes clústeres). Pueden existir otro tipo de representaciones de estos mapas (Vensanto, et. al., 2000).

3- Mapas de planos.- Son la representación de los mapas de distancia entre vecinos (todas las componentes), pero para cada una de las variables o componentes del espacio de entrada. Muestran cuales variables tienen mayor influencia en el resultado del clustering (Stefanoic y Kurasova, 2011). Los colores oscuros (Fig. 3.14) representan pesos más grandes, las variables de entrada con gradientes de colores similares (o inversos), pueden representar alta correlación (o falta de ella). La información proporcionada es muy valiosa y útil, en los casos de espacios de entrada multicomponente, donde las representaciones gráficas se vuelven complicadas.

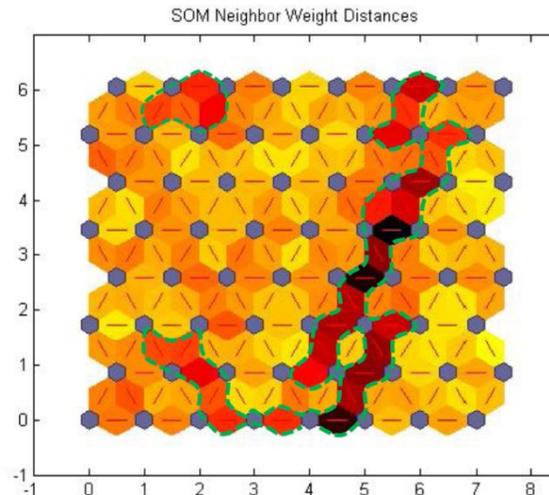


Fig. 3.13 Ejemplos Mapa de distancia entre vecinos, Manual de Matlab (2009)

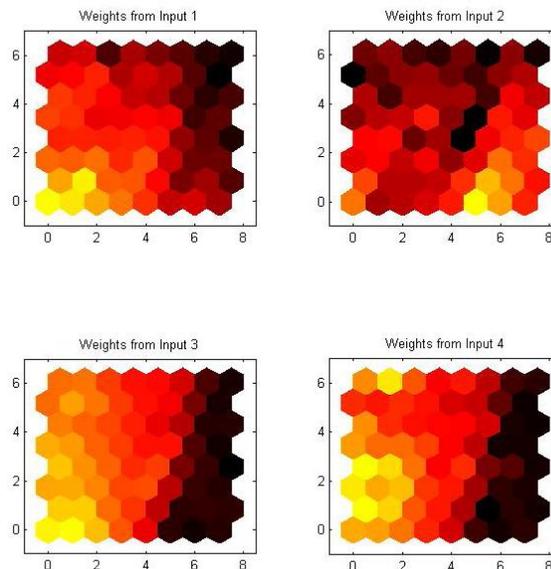


Fig. 3.14 Ejemplos de mapas de planos, para cuatro entradas, Manual de Matlab (2009)

3.7 Clasificación multi-pozo de electrofacies mediante el modelo neuronal.

Una vez entrenados los modelos neuronales, con un solo pozo, se procede a simular la respuesta de salida de la red para los demás pozos. La elección del pozo que entrenará la red dependerá de la estadística de los datos, la consistencia de la clasificación o guía (patrón de salidas), la distribución espacial de los demás pozos y el modelo de depósito (los pozos debe haber cortado las mismas electrofacies). El contar con información adicional como núcleos, sísmica, estudios paleontológicos, permiten tener claro el modelo geológico conceptual de un área, que permitirá tomar la decisión de que pozo es el mejor para entrenar la red. Lo anterior es clave para definir el patrón de clasificación, que será el guía o maestro durante el entrenamiento de los modelos supervisados, ya que de ello dependerá el resultado de la clasificación de los demás pozos.

En el caso de los modelos no supervisados, el conocimiento de la geología del área, nos permite interpretar el número de clases obtenidas y los datos de núcleo o interpretaciones manuales, nos ayudan a validar y calibrar nuestros modelos.

Para este trabajo de tesis se propone el siguiente esquema para realizar la estimación de electrofacies:

1- Para el entrenamiento y prueba, se escogerán los pozos que sea estadísticamente más representativo de las electrofacies (que se esperan hayan cortados las mismas litofacies y tengan la misma información de entrada).

2- En el caso de modelos supervisados, se escoge un pozo para entrenar el modelo neuronal, con el que se estimarán los demás pozos.

3- En el caso de los modelos no supervisados se presentan dos opciones: a) Se entrena un *SOM* con un solo pozo para estimar los demás pozos o b) se entrena el *SOM* con cada uno de los pozos de forma individualmente. En este trabajo se propone la opción 1, ya que entrenar la red con los datos de cada pozo, limita el proceso de validación cruzada.

4- Para escoger el pozo de entrenamiento, se realizarán diversas pruebas con las redes propuestas. En el caso de los modelos supervisados, se deberán escoger pozos que tengan una clasificación bien definida (realizada por algún especialista y de preferencia corroborada con datos de núcleo). Lo anterior también es altamente recomendable en los modelos no supervisados, ya que esto permite validar los resultados.

5- Una vez generado el modelo neuronal y entrenado con el pozo seleccionado, se procede a estimar las electrofacies de los demás pozos.

6- Si los pozos candidatos a la estimación de electrofacies, tienen una clasificación previa, se procede a realizar el análisis del error mediante el análisis de las matrices de confusión

7- Interpretar y analizar los resultados.

IV. ANÁLISIS DE LA APLICACIÓN DE *REDES NEURONALES ARTIFICIALES* A DATOS SINTÉTICOS

Como se mencionó en el capítulo anterior, el uso de registros sintéticos permite realizar las pruebas necesarias para identificar parámetros claves durante el diseño y entrenamiento de la red neuronal. Generalmente con estas pruebas se pretenden determinar los aspectos asociados con la arquitectura de la red.

En el presente capítulo se muestran los resultados obtenidos al aplicar la metodología para estimación automática de electrofacies propuesta en este trabajo, a datos sintéticos de registros de pozo, persiguiendo los siguientes objetivos:

1. Probar el flujo de trabajo en condiciones controladas. Para ello se proponen tres modelos definidos por el grado de clasificación entre los datos (Bien clasificado, medianamente clasificado y mal clasificado), así como el nivel de ruido aleatorio, que contaminan los datos durante la clasificación de electrofacies.
- 2- Comprender de mejor forma los resultados obtenidos en cada una de las etapas de esta metodología, e inferir la estrecha relación que existe entre la naturaleza estadística del dato y los resultados de la clasificación mediante las redes neuronales.
- 3- Definir parámetros importantes como el número de capas, el número de neuronas y en su caso el número de redes modulares a ejecutarse (ver sección 3.5.6.3).
- 4- Evaluar el rendimiento del entrenamiento de la red neuronal, para inferir el rango en el que los parámetros de cada tipo de modelo neuronal, se deberían encontrar.
- 5- Evaluar los resultados y el error del pronóstico de electrofacies estimadas a partir de los registros sintéticos y sacar conclusiones antes de aplicar la metodología a un caso real de estudio.

4.1 Registros geofísicos y criterio de clasificación

Como se mencionó en el capítulo uno el término facie tiene diferentes concepciones y su aplicación como criterio de clasificación o zonificación de los registros de pozo, está en función de los fines que se persigan. Como se mencionó anteriormente (capítulo 1), el objetivo de la clasificación propuesta en este trabajo de tesis, es con fines de caracterización de propiedades petrofísicas en un yacimiento, dentro del marco de la Modelación Geológica Petrofísica de Yacimientos, teniendo como criterio de clasificación en el contenido de arcillosidad estimado del registro de rayos gama (pseudo-litológico).

Para el desarrollo del presente capítulo, se generan registros geofísicos sintéticos a partir de modelos de secuencias areno-arcillas, que simulan algunos de los patrones de depósito del modelo de abanico de la Fig. 4.1. Cada secuencia (arenas limpias, lutitas ó arenas arcillosas laminares o no), están caracterizada por propiedades petrofísicas específicas (fluidos, porosidad, volumen de arcilla, etc.), con el objetivo de analizar los resultados obtenidos de la clasificación de registros con redes neuronales, aplicadas a secuencias litológicas similares (arenas y/o lutitas) con diferentes respuesta en los registros de pozo (Fig. 4.2). En la Fig. 4.2 se muestran de manera esquemática, las respuestas típicas de los registros de pozo convencionales para diferentes litologías y diferente contenido de fluido.

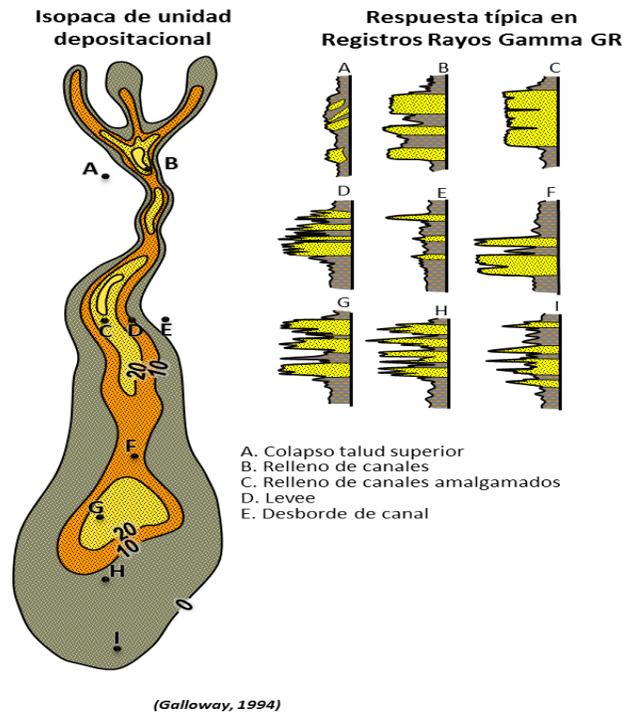


Fig. 4.1 Patrones de apilamiento en la respuesta de los registros de rayos gama (electrofacies) asociadas a facies de depósito en un modelo de abanico (Galloway, 1998)

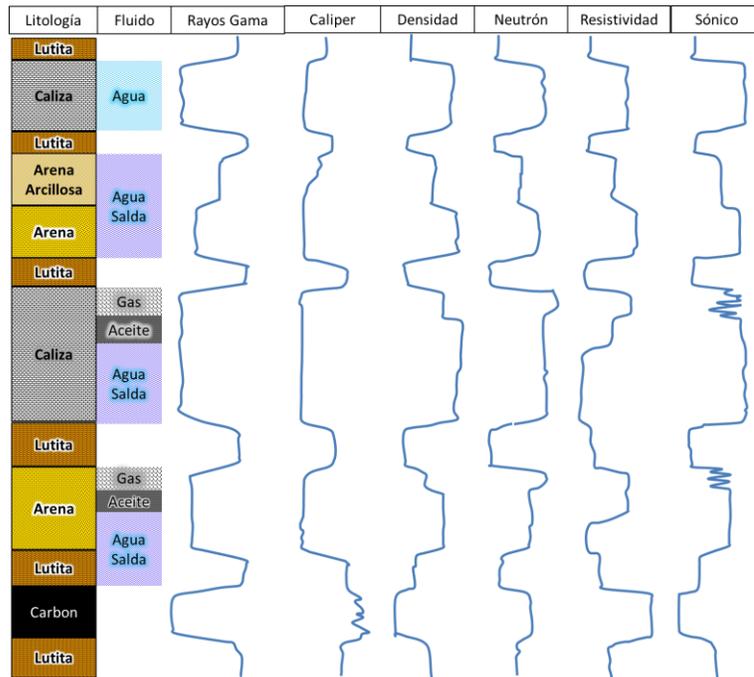


Fig. 4.2 Respuestas típicas de los registros convencionales de pozos, en litologías y tipos de fluidos específicos (Arm Geophysics, 2013)

4.2 Registros sintéticos

Para la construcción de los registros de pozo sintéticos se consideraron tres modelos, que representan tres escenarios controlados:

- a) Buena clasificación en la respuesta de los registros.- Existe una diferencia obvia entre las respuestas de los registros sintéticos asociados a las clases propuestas.
- b) Regular la clasificación en la respuesta de los registros.- La identificación de las facies propuestas, es menos clara en las respuestas de los registros sintéticos.
- c) Mala clasificación en la respuesta de los registros.- Las respuestas de los registros sintéticos, no permite identificar muy claramente, las diferentes facies propuestas.

4.3 Generación de las curvas sintéticas

Para la generación de los registros sintéticos se consideraron tres facies o patrones de depósito de sistemas de abanicos submarinos de la Fig. 4.1:

- a) Facie A.- Facies de canal (Fig. 4.1 C)
- b) Facie B.- Lutitas francas
- c) Facie C.- Facies de desbordes de canal o leves (Fig. 4.1 D)

Las curvas de los registros sintéticos se generaron mediante expresiones petrofísicas teóricas, a partir de modelos de secuencias de arenas, arenas arcillosas ó lutitas, con propiedades petrofísicas diferentes (de matriz y fluidos). Los valores obtenidos, se variaron en un rango de valores específicos para cada facie, obtenidos de registros de campos análogos (Jiménez y Sandoval, 2009), tratando de simular los patrones de depósito de la Fig. 4.1 C y D. Es importante mencionar se introduce un factor de ruido aleatorio, de 5, 10 y 15 por ciento, para el modelo uno, dos y tres respectivamente.

4.3.1 Curva de rayos gama.

El registro de rayos gama responde al contenido de minerales radioactivos como potasio, torio y uranio, los cuales se encuentran principalmente asociados a minerales arcillosos, por lo que el registro es un buen indicador del contenido de arcilla. La Ec. 4.1, representa la expresión típica para obtener el volumen de arcilla a partir del registro de rayos gama.

$$V_{sh} = \frac{GR_{registro} - GR_{min}}{GR_{max} - GR_{min}} \quad 4.1$$

Estableciendo los parámetros de volumen de arcilla (V_{sh}) y los valores de rayos gama máximo (GR_{max} para lutitas francas) y mínimo (GR_{min} para arenas francas) de la literatura (Crain, 2013) y de trabajos publicados (Jiménez y Sandoval, 2009) de campos como: Spraberry (Texas, E.U.), Priobskoye (Siberia, Rusia) y Xinzhaio (Daqing, China).

4.3.2 Curva de resistividad.

El registro de resistividad responde principalmente al contenido y tipo de fluidos, contenidos en el espacio poroso, por lo que sus valores responden a diferentes parámetros tanto de la matriz, la mineralogía, el cementante, la compactación, el tipo de porosidad, la arcillosidad, la salinidad del agua de formación, el tipo de hidrocarburo, etc. El registro sintético se determinó utilizando las siguientes expresiones de las ecuaciones 4.2 (ecuaciones de Archie de 1942), para arenas limpias y las ecuaciones 4.3 (Bardon y Pied, Modificadas por Simandoux en 1969) y 3.4 (Crain, 2013), para arenas arcillosas.

$$R_t = \frac{aR_t}{\phi^m S_w^n} \quad 4.2$$

$$\frac{1}{R_t} = \frac{S_w^2}{FR_w} + \frac{V_{sh}S_w}{R_{sh}} \quad 4.3$$

$$V_{sh} = \frac{\log(R_t) - \log(R_{Arena\ limpia})}{\log(R_{sh}) - \log(R_{Arena\ limpia})} \quad 4.4$$

Fijando para cada facie propuesta, los valores de saturación de agua (S_w), volumen de arcilla (V_{sh}) y porosidad (ϕ) y proponiendo los exponentes a , m , n , resistividad en una arena limpia ($R_{Arena\ limpia}$) y la resistividad de una lutita franca (R_{sh}), de la bibliografía antes citada.

4.3.3 Curva Caliper

La curva del registro caliper, responde principalmente a derrumbes y zonas permeables, donde el enjarre del lodo (capa que se forma al filtrarse el lodo de perforación en la formación). En zonas arcillosas y zonas no consolidadas, los derrumbes son frecuentes, por lo que un indicador de una zona arcillosa puede ser el registro caliper. El registro sintético del caliper se determinó al variar los valores observados en la bibliografía de los campos análogos, en un rango determinado, enfatizándose en las zonas arcillosas.

4.3.4 Curva de densidad

El registro de densidad proporciona una estimación de la densidad volumétrica (ρ_b) de un cuerpo rocoso, Dicho parámetro depende de la densidad de la matriz (ρ_{ma}) y la densidad de los fluidos (ρ_{fl}). Las expresiones utilizadas para el cálculo de densidad fueron consultadas en Dresser Industries (1979)

Para la densidad de la matriz se utilizó la Ec. 4.5, definiendo el volumen de arcilla (V_{sh}) la densidad de la arena (ρ_{arena}) y la densidad de las lutitas (ρ_{sh}), de la bibliografía de los campos análogos.

$$\rho_{ma} = (1 - V_{sh})\rho_{arena} + V_{sh}\rho_{sh} \quad 4.5$$

Para la densidad del fluido se utiliza la Ec. 4.6, fijando los parámetros saturación de agua (S_w), saturación de aceite (S_o) y saturación de gas (S_g), para cada una de las facies propuestas, además se utilizaron valores de densidad del agua, densidad del gas y densidad del aceite de Alpana (2002).

$$\rho_{fl} = S_w\rho_w + S_o\rho_o + S_g\rho_g \quad 4.6$$

El cálculo del valor de densidad en función del volumen de arcilla (V_{sh}) y del espacio poroso efectivo (ϕ_e) viene dada por la Ec. 4.7

$$\phi_e = \frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_{fl}} - V_{sh} \frac{\rho_{ma} - \rho_{sh}}{\rho_{ma} - \rho_{fl}} \quad 4.7$$

4.3.5 Curva de neutrón

El registro de neutrones responde principalmente al contenido de hidrogeno, presente en los fluidos de la roca. Para el registro de neutrones sintético, se utilizaron valores medios obtenidos de los campos análogos citados en la bibliografía y comparando con la Ec. 4.8 encontrada en Schlumberger (1975).

$$\phi_{N-sh} = \phi_N - 0.3V_{sh} \frac{\phi_{Nsh}}{0.45} \quad 4.8$$

Dónde:

ϕ_N = porosidad neutron

ϕ_{N-sh} = porosidad neutrón corregida por arcillas

V_{sh} = volumen de arcilla

ϕ_{Nsh} = valor de la porosidad neutrón medida en una arcilla

4.3.6 Curva tiempo de tránsito (sónico)

El registro sónico mide los tiempos de arribo de ondas compresionales y de cizalla, que atraviesan la formación, por lo que la respuesta de este tipo de registros responde tanto a matriz de roca como a fluido contenido en ella.

Para estimar el registro sónico sintético se determinaron el tiempo de tránsito de la matriz (Δt_{ma}) de la Ec. 4.9 (Alpana, 2002), fijando los valores del volumen de arcilla (V_{sh}) y los tiempos de tránsito medidos en las lutitas (Δt_{sh}) y las arenas limpias (Δt_{arena}) de los campos análogos y valores de velocidad de bibliografía sísmica (Jones, 2010).

$$\Delta t_{ma} = (1 - V_{sh})\Delta t_{arena} + V_{sh}\Delta t_{sh} \quad 4.9$$

Para estimar los valores de tiempo de tránsito del fluido, se utilizó la Ec. 4.11 y los valores mostrados en Alpana (2002).

$$\Delta t_{fl} = S_w\Delta t_w + S_o\Delta t_o + S_g\Delta t_g \quad 4.10$$

Para estimar los tiempos de tránsito de las arenas limpias y arenas arcillosas, se utilizaron las ecuaciones 4.11 (Wyllie et. al, 1956 y 1958) y 4.12 (Dresser Industries, 1979) respectivamente.

$$\phi = \frac{\Delta t_{log} - \Delta t_{ma}}{\Delta t_{ma} - \Delta t_{fl}} \quad 4.11$$

$$\phi = \left[\frac{100}{\Delta t_{sh}} * \frac{\Delta t_{log} - \Delta t_{ma}}{\Delta t_{fl} - \Delta t_{ma}} \right] - V_{sh} \left[\frac{\Delta t_{sh} - \Delta t_{ma}}{\Delta t_{fl} - \Delta t_{ma}} \right] \quad 4.12$$

4.3.7 Litotipos

En este trabajo los litotipos (litofacies) son el objetivo principal de clasificación, mediante la interpretación de electrofacies. Los litotipos seleccionados para la construcción de modelos son modelos siliciclásticos de arenas y lutitas. Las electrofacies serán definidas principalmente de las respuestas de los registros de pozo a propiedades petrofísicas del tipo de roca con porosidad, volumen de arcilla y el contenido de fluidos, específicos. Los conjuntos de clases petrofísicas o electrofacies estimadas, serán nuevamente clasificados e interpretados como facies de depósito.

Es importante resaltar que las respuestas de los registros geofísicos de pozo pueden variar, dentro de un mismo litotipo, si las condiciones porosidad, permeabilidad, arcillosidad y contenido de fluidos es diferente.

4.3.8 Modelo uno (Bien clasificado)

El modelo uno, simula las respuestas de tres secuencias arenosas y cuatro secuencias arcillosas, que representan las facies arcillosas más distales del abanico, depósitos de arenas de canal y las facies areno-arcillosas de desborde de canal (Fig. 4.1). La Fig. 4.4 muestra el modelo y los parámetros utilizados para generar los registros sintéticos graficados en la Fig. 4.5.

El modelo uno representa el caso bien clasificado, basados principalmente en las respuestas de los registros rayos gama y resistividad. Sin embargo puede notarse en la Fig. 4.5, que en menor medida, los registros de densidad, neutrón y sónico también permiten zonificar este set de registros.

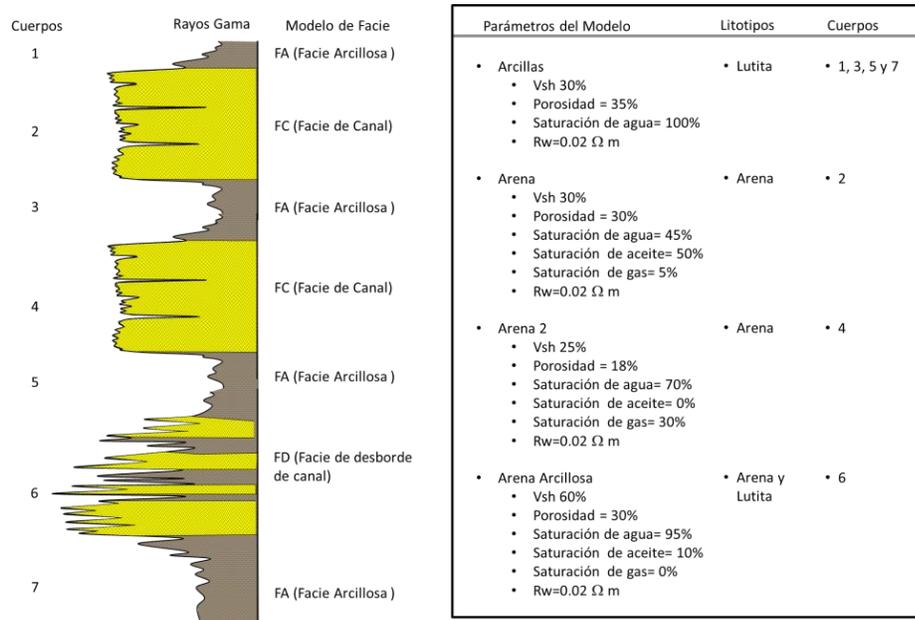


Fig. 4.3. Modelo uno

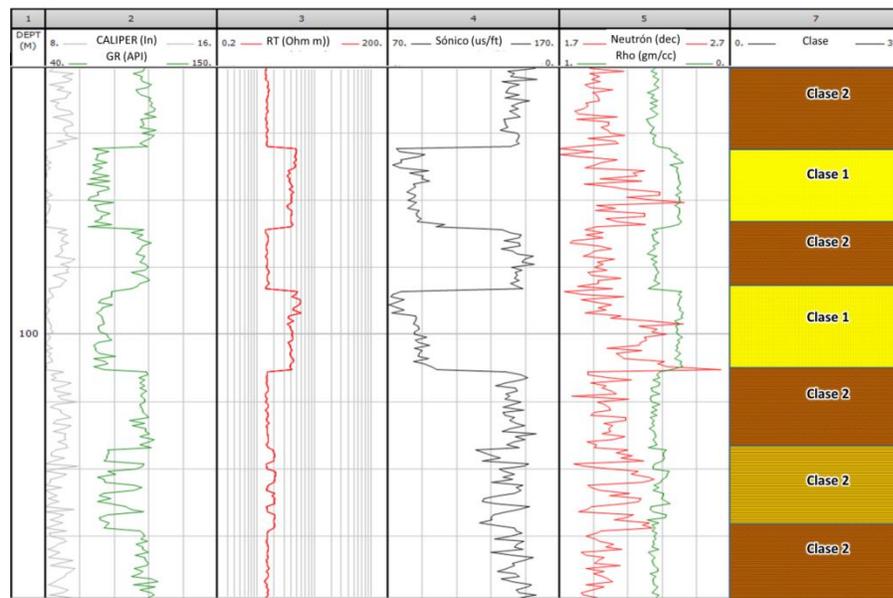


Fig. 4.4 Registros geofísicos de Pozo sintéticos del Modelo uno.

4.3.9 Modelo dos (Medianamente clasificado)

El modelo dos, simula las respuestas típicas del registro de rayos gama de dos secuencias de facies de depósitos de arenas de canal y facies areno-arcillosas de desborde de canal; en la Fig. 4.5 muestra los parámetros del modelo y en la Fig. 4.5 registros sintéticos.

El modelo dos representa el caso bien medianamente clasificado. En la Fig. 4.6 puede notarse que el contraste entre los registros es menor y diferente para todos los registros, principalmente el registro sísmico

4.3.4 Modelo tres (Mal clasificado)

El modelo tres, muestra las respuestas del registro de rayos gama de tres secuencias de facies de depósitos de arenas de canal y facies areno-arcillosas de desborde de canal. Los registros sintéticos generados para este modelo se ilustran en la Fig. 4.7.

El modelo tres representa el caso mal clasificado. En la Fig. 4.8 puede notarse que el contraste entre los registros es el mas mala en todos los modelos y solo en unos cuantos registros pueden diferenciarse la zonificación, principalmente atribuida a la respuesta de las facies de arenas de canal.

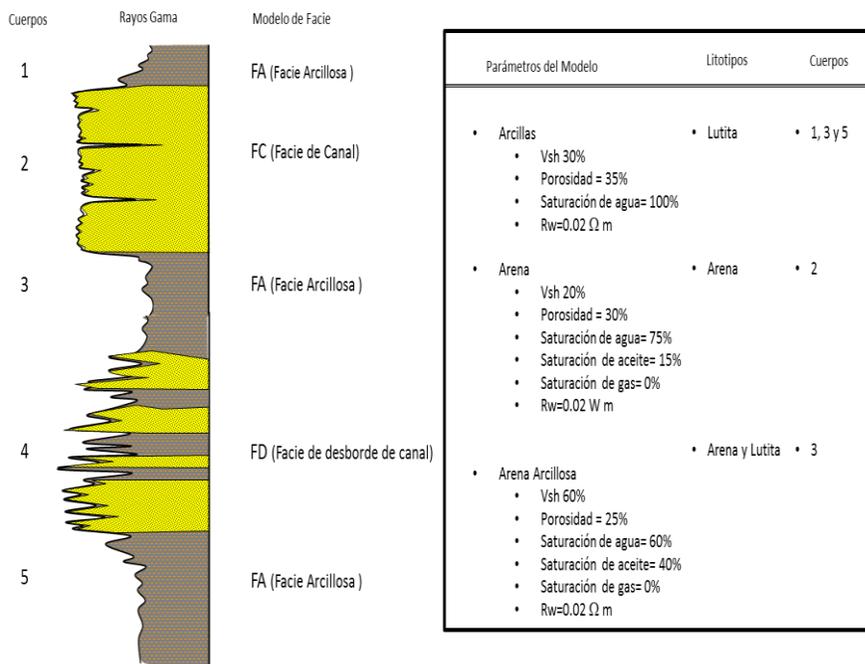


Fig. 4.5. Modelo dos

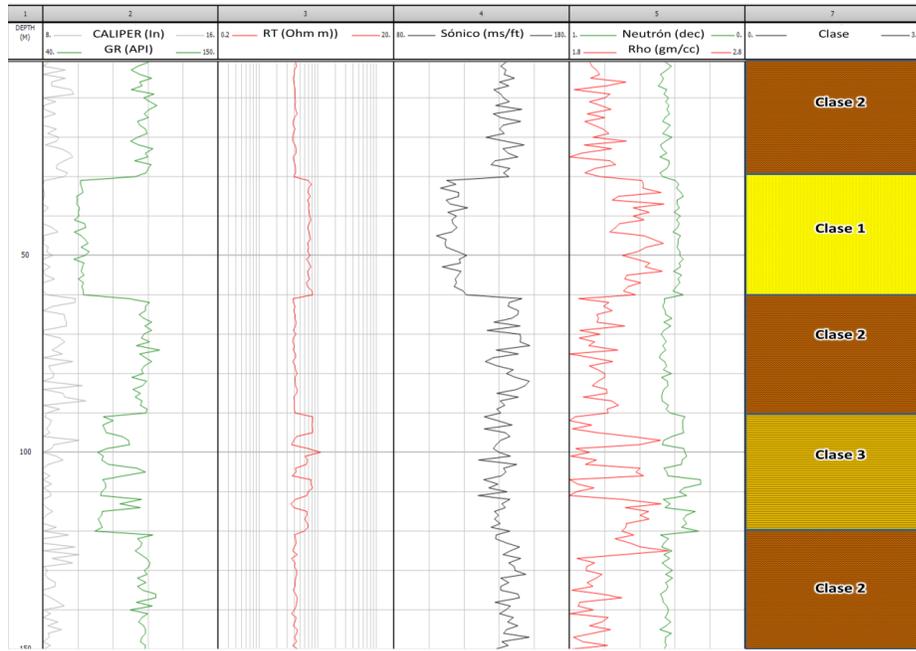


Fig. 4.6. Registros geofísicos de Pozo sintéticos del Modelo dos.

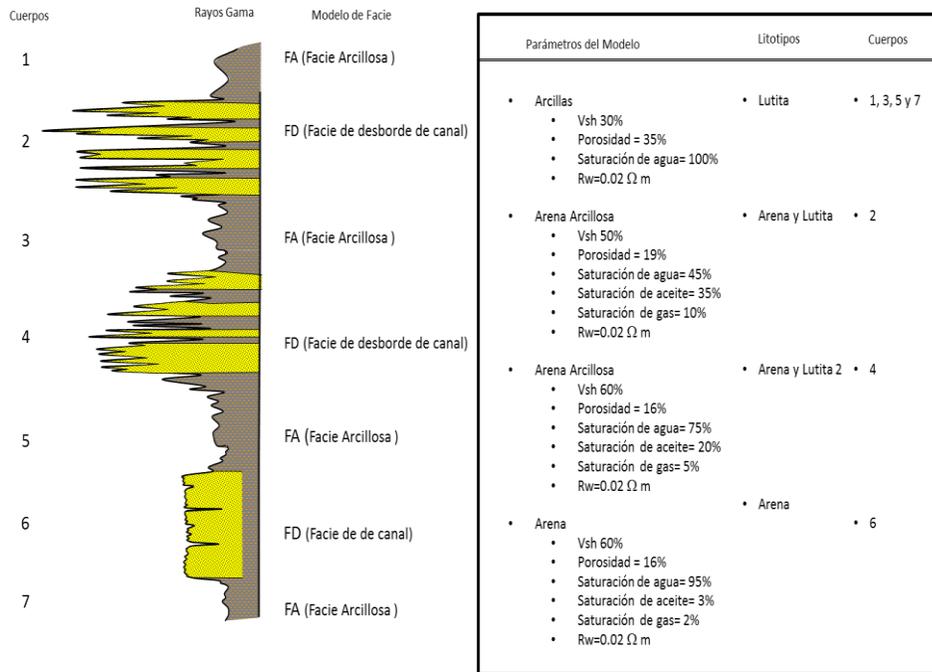


Fig. 4.7. Modelo tres

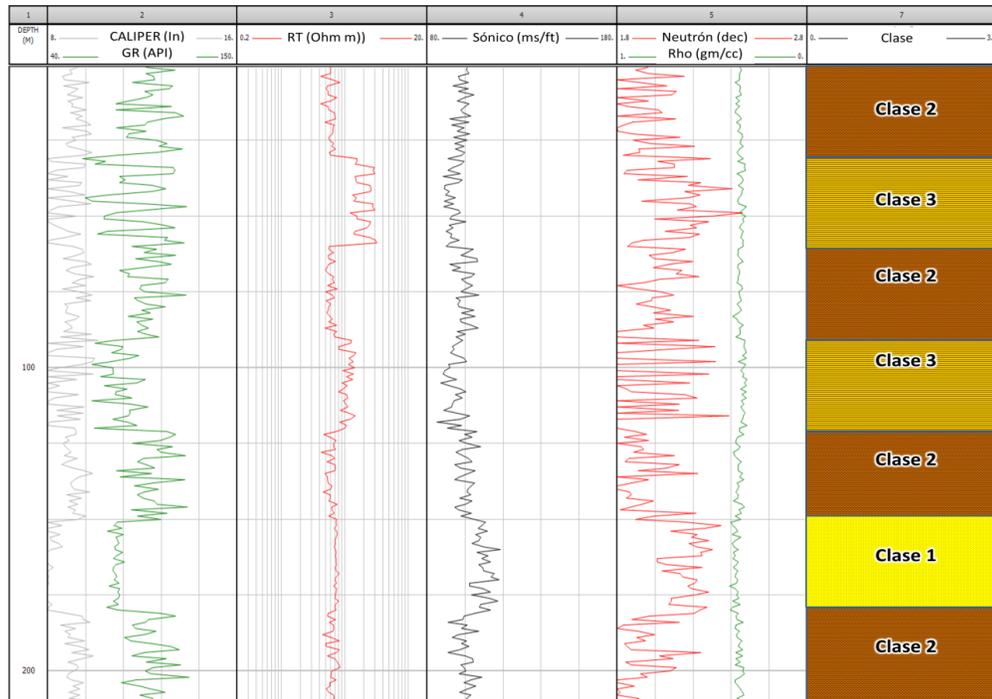


Fig. 4.8. Registros geofísicos de Pozo sintéticos del Modelo tres.

4.4.1 Censo y revisión de la información (control de calidad).

Esta etapa de la metodología se obvia, ya que los registros son sintéticos, sin embargo es importante analizar la calidad y veracidad del dato, con algún petrofísico, lo importante es tomar en revisar que los datos correspondan al registro e intervalo en estudio, la falta de información (puntos sin datos) y posibles datos anómalos basados únicamente en la experiencia de algún experto. El propósito de lo anterior es construir los archivos, que serán los insumos necesarios para comenzar a realizar los cálculos y análisis de este flujo de trabajo.

4.4.2 Análisis exploratorio de datos

El análisis exploratorio de datos se realizó en su totalidad con el lenguaje de programación R. El objetivo es analizar el comportamiento estadístico de los datos de los tres modelos. Para realizar este análisis se realizan:

- El cálculo de estadísticos básicos y los gráficos estadísticos, por clase.
- El análisis estadístico de cada uno de los modelos de manera individual
- El cálculo de estadísticos básicos y la generación de gráficos estadísticos, por modelo y clase.

Los resultados del análisis exploratorio de datos (Tablas y gráficos), se ilustran en el apéndice A, de este trabajo de tesis.

4.4.2.1 Análisis por clase

Las Tablas A.1, A.2 y A.3, del apéndice A, muestran los estadísticos de los tres modelos propuestos (uno, dos y tres), por clase, permiten analizar los valores de los registros que caracterizan a cada clase y

ver las diferencias o similitudes entre clases, para un registro en específico. Por ejemplo la clase uno es contrastante de las clases dos y tres considerando la media del registro de rayos gama (Hay que recordar que la clase dos es una lutita franca y la clase tres es una facie areno-arcillosa). Otro ejemplo sería los valores medios del registro de resistividad, donde se aprecian bajos contrastes; en cambio, la relación RT/GR muestra mejor diferenciación entre los valores medios de cada clase.

Del análisis de estas mismas Tablas (A.1, A.2 y A.3), se observa que los registros que contrastan mejor para caracterizar cada una de las tres facies (utilizando de criterio la mediana) se encuentra el registro RT/GR , sónico, neutrón y caliper. También se puede analizar, el grado de variabilidad, de los datos para cada clase. Donde la clase uno y dos tienen menor varianza (ya que se tratan de arenas más limpias y lutitas, respectivamente), en cambio, la clase tres varía en un rango mayor de valores prácticamente en todos los registros, lo que permite suponer que algudatos de esta clase, tienen valores en los rangos de las clases uno y dos.

El análisis de variabilidad en cada clase, da información sobre:

- 1) Datos atípicos univariados (outliers)
- 2) La presencia de subclases desde el punto de vista univariado, que no están suficientemente muestreadas (alternancia de diferentes litotipos, como en el caso de la clase tres y/o diferente contenido de fluidos en una misma clase)
- 3) Un efecto combinado

Las figuras A.1, A.2 y A.3, del apéndice A, muestran los histogramas y los gráficos de cajas de los registros para cada clase. En estas imágenes, se observan las simetrías de los histogramas y la presencia de probables valores atípicos. Lo esperado sería tener una distribución uniforme de los datos ya que se esperaría que cada clase fuera caracterizada por la media o la mediana y por un rango de valores muy cercano a ellas. Sin embargo se pueden observar distribuciones sesgadas (Ejemplo el registro de resistividad, la relación RT/GR) y algunas con posibles modas (Rho, sónico, neutrón).

En este caso la asimetría puede deberse a que los registros sintéticos fueron generados a partir de tres modelos con diferente grado de clasificación, por lo que la simetría de algunas variables, pueden estar diferenciando cada uno de los modelos, lo cual implica la falta de consistencia en la clasificación (se debería replantear el criterio de clasificación, por ejemplo la clase tres podría ser solo una alternancia de clase uno y dos). La clase dos fue la más consistente ya que para los tres modelos se usaron prácticamente los mismos parámetros.

Para intentar mejorar la distribución de los datos, se analizaron los datos sin valores atípicos y aplicándoles una transformación logarítmica, ya que la escala del registro de resistividad generalmente es logarítmica (estos análisis fueron realizados más no se muestran en el apéndice A).

De este último análisis se puede concluir que:

- 1) Remover datos atípicos, mejoraba la distribución de los datos, pero implicaba la pérdida de información valiosa.
- 2) Utilizar transformaciones como la logarítmica incrementaba el nivel de ruido en algunas de las variables
- 3) Eliminando los valores atípicos y aplicando la transformación logarítmica, mejora solo en algunos casos la distribución de los datos (GR , Caliper y densidad) y en otras se enfatizaba las diferentes modas (Neutrón, sónico y resistividad). Las diferentes modas se pueden interpretar como la presencia de subclases.

- 4) La clase dos presenta una mejor distribución, seguido de la clase uno y posteriormente la clase tres (mezcla de litotipos de arena y arcilla).

Las figuras A.4, A.5 y A.6, muestra los gráficos cuantil - cuantil, donde se puede observar cualitativamente, el grado de aproximación de una distribución normal. En general para todas las clases los registros de rayos gama, densidad, sónico y caliper, son los que mejor se aproximan a una distribución normal. La Fig. A.7 muestra los gráfico de cajas por clase, para cada una de las variables, en donde se observa lo explicado en las Tablas A.1, A.2 y A.3, Siendo los registros de rayos gama, resistividad, caliper, sónico, neutrón y la relación RT/GR , lo que más diferencian las clases con respecto al valor medio, sin embargo, en todos los caso existe traslape de los rangos intercuartílicos, lo que da idea de mala clasificación en general, desde el punto de vista univariado.

La Fig. A.8 del apéndice A, muestra los gráficos de dispersión (o scatterplots), en donde se observa el grado de correlación (coeficiente de correlación de Pearson) entre variables y el conjunto de subclases bivariadas existentes. De esta figura se puede ver marcada la tendencia de cada uno de los modelos, además de los pares de registros que de forma bivariada responden más a esta clasificación.

4.4.2.2 Análisis por modelo y por modelo-clase

El análisis por modelo y por modelo-clase permite conocer el comportamiento global de los datos, así como visualizar si para cada variable (o pares de ellas), por sí mismas, son capaces de diferenciar cada una de las clases, en cada modelo.

La Fig. A.9 muestra la semejanza de los modelos a través de los valores medios representados en los gráficos de cajas, de donde se puede concluir de manera muy general, que los modelos uno y dos, son más parecidos con respecto a la media, a diferencia del modelo tres que pareciese tener un comportamiento distinto a los otros dos modelos. No obstante lo anterior, los rangos intercuartílicos son muy diferentes, este efecto es asociado al número de datos que pertenecen a una clase u otra.

a) Modelo uno

En la Tabla A.4 del apéndice A se muestran los estadísticos generales de las variables en estudio para el modelo uno. La información mostrada permite inferir el grado de variabilidad de los datos y da una idea de cómo será la distribución de los mismos.

La Fig. A.10, muestra los histogramas y gráfico de cajas de los registros del modelo uno, se puede observar que las distribución de los datos en los histogramas para los registros GR , RT , sónico, neutrón y la relación RT/GR , son los que muestran mejor respuesta a las diferentes clases, debido a que presentan diferentes modas. A diferencia de los registros densidad y caliper (principalmente zonas arcillosas) y la derivada de GR con respecto a la profundidad, muestran una distribución homogénea independiente de la clase.

La Fig. A.14, muestra los gráficos cuantil- cuantil, donde se observa que precisamente la distribución de los registros de densidad y caliper, se aproximan mejor a una distribución normal. La derivada de los rayos gama con respecto a la profundidad se aproxima en una parte y en los extremos se separa, siendo esos puntos los valores atípicos mostrados por los gráfico de cajas de la Fig. A.10.

La Fig. A.15, muestra los gráficos de dispersión (scatterplots) y los coeficientes de correlación entre las variables (o pares de variables), donde se observa el grado de clasificación bivariada de este modelo (Bien clasificado), en donde la relación resistividad- rayos gama, es que la que mejor define los grupos prácticamente con todos los demás registros, seguido del neutrón, rayos gama y resistividad. Se observa también que para ciertos pares de variables, existe mala clasificación (por ejemplo densidad y caliper o caliper y rayos gama), sobre todo para las clases uno y tres.

Las Tablas A5, A6 y A7, permiten ver el rango de valores en el que se encuentran los datos de los registros para cada clase (uno, dos y tres) respectivamente, en el modelo uno. Las figuras A.11, A.12 y A.13, muestran los histogramas de los datos del modelo uno, para cada clase. De este análisis se puede observar que evidentemente la clase uno (Arenas más limpias) y las lutias, presentan una mejor distribución, por lo que sus valores varían menos con respecto al valor medio que los caracteriza. Sin embargo, la clase tres presenta distribuciones de los datos con diferentes modas, debido a que esta clase es una combinación de litotipos (arenas y lutitas). Los gráficos cuantil-cuantil por modelo y clase, no se muestran en este trabajo, sin embargo las variables con histogramas más simétricos son los que más se aproximan a la recta de distribución normal del gráfico cuantil- cuantil.

b) Modelo dos

En la Tabla A.8 del apéndice A se muestran los estadísticos generales de las variables en estudio para el modelo dos (medianamente clasificado). La información mostrada ahí y los histogramas de la Fig. A.16, permite inferir el grado de variabilidad de los datos y da una idea de cómo es la distribución de los mismos. Para este análisis los registros rayos gama, el registro de resistividad, el registro sínico, el neutrón y la relación resistividad - rayos gama, muestran multimodalidad y probablemente sean los datos que mejor representen las tres clases definidas.

La Fig. A.20, muestra los gráficos cuantil- cuantil, donde se observa al igual que el modelo uno, distribución de los registros de densidad, caliper y la derivada de los rayos gama con respecto de la profundidad, se aproximan mejor a una distribución normal.

La fig. A.21, muestra los gráficos de dispersión y los coeficientes de correlación entre las variables. Nuevamente se observa buena clasificación de los datos, principalmente, en los siguientes pares de registros:

- a) Rayos gama con resistividad, caliper, densidad, sínico y neutrón
- b) Resistividad con sínico
- c) Sínico con neutrón y relación RT/GR
- e) Relación RT/GR con neutrón, densidad y caliper.

Las Tablas A9, A10 y A11, muestran los estadísticos de este modelo para las clases uno, dos y tres. Las figuras A.17, A.18 y A.19, muestran los histogramas de los datos del modelo dos analizados por clase. Los análisis por clase, permiten observar que el comportamiento de los registros es muy similar, para cada clase entre el modelo uno y el modelo dos, sin embargo los rangos de valores son diferentes, las varianzas tienden a ser similares, por lo que una primera hipótesis es que la clasificación en ambos modelos es muy similar (por ejemplo la clase uno del modelo uno es similar, pero no igual a la clase uno del modelo dos). Nuevamente la clase tres es la que presentaría un número de subclases univariadas, debido a que esta clasificación es asociada a dos litotipos diferentes.

Al igual que en el modelo uno, los gráficos cuantil - cuantil por modelo y clase, no se muestran en este trabajo.

c) Modelo tres

La Tabla A.12, muestra los estadísticos del modelo tres (relativamente mal clasificado), La información mostrada ahí y los histogramas de la Fig. A.22, muestran que los datos presentan una mejor distribución, debido a que los datos se encuentran mal clasificados. Solo los registros de resistividad, caliper y la relación RT/GR , muestran diferencias en los datos.

La Fig. A.30, muestra los gráficos cuantil- cuantil, donde la mayoría de los registros (a excepción del registro de resistividad, el registro del caliper y la relación RT/GR) se aproximan en la mayoría de los puntos a una tendencia próxima a una distribución normal, esto debido a la mala clasificación de los datos, ya que se ven afectados por el ruido aleatorio.

La Fig. A.27, muestra los gráficos de dispersión y los coeficientes de correlación entre las variables. Donde prácticamente todos los registros muestran una mala clasificación y bajos coeficientes de correlación. Nuevamente los registros de resistividad y la relación RT/GR , son los que presentan mejor clasificación. La clasificación es variada entre los diferentes pares de registros, por ejemplo en el par resistividad y densidad, las clases tres y dos, son las que más se parecen. Y en el par rayos gama y sónico la clase dos y uno, son las más parecidas, pero en general la clase arcillosa, es la que presenta mayor dispersión, prácticamente en todos los registros.

Las Tablas A13, A14 y A15, muestran los estadísticos de este modelo para las clases uno, dos y tres. Las figuras A.23, A.24 y A.25, muestran los histogramas de los datos del modelo tres analizados por clase. En esta ocasión se observa que el comportamiento de los registros es diferente al mostrado en los modelos uno y dos, ya que la forma y simetría de los histogramas, permiten inferir diferentes subclases (a excepción de la clase dos, la cual es una lutita franca y tiene distribuciones de los datos, más homogéneas), los rangos de valores y las varianzas, son más grandes que los modelos uno y dos, por lo que se puede inferir que el grado de ruido que pudieran presentar los datos, está impactando en los valores que caracterizan cada clase.

4.4.2.3 Conclusiones del análisis exploratorio de datos.

Del análisis exploratorio, se puede concluir que los diferentes modelos cumplen su cometido, en cuanto al grado de clasificación, también se puede observar que un sistema de clasificación tendría mayor problema al tratar de clasificar el modelo tres. Prácticamente la clasificación en el modelo tres no es consistente con las del modelo uno y dos, por lo que se esperaría como hipótesis, resultados con mayor incertidumbre, al tratar de clasificar el modelo tres.

Con este análisis se puede ver el rango de valores en los que se encuentra cada facie aproximadamente y los registros que tenderán a discriminar mejor las clases. En este contexto los registros que permitieron mejor la discriminación entre clases fueron la relación RT/GR , la resistividad, los rayos gama, el neutrón y el sónico.

De manera gráfica y cualitativa, se observaron grupos o clústeres bivariados, dando un panorama general previo al Análisis de agrupamiento. El análisis de correlación lineal entre los diferentes registros, da información valiosa, previa al análisis de componentes principales y la selección de las variables más óptimas para el entrenamiento.

4.4.3 Análisis de agrupamiento

El Análisis de agrupamiento se realizó en su totalidad por medio del lenguaje R.

Como se explicó en el capítulo dos, lo importante del Análisis de agrupamiento, es establecer un criterio de similaridad, el cual, después de algunas pruebas se determinó utilizar la distancia euclidiana, así como el algoritmo no jerárquico de las k-medias.

a) Análisis de Dendrogramas

Una práctica común es realizar un pre-análisis jerárquico utilizando dendrogramas.

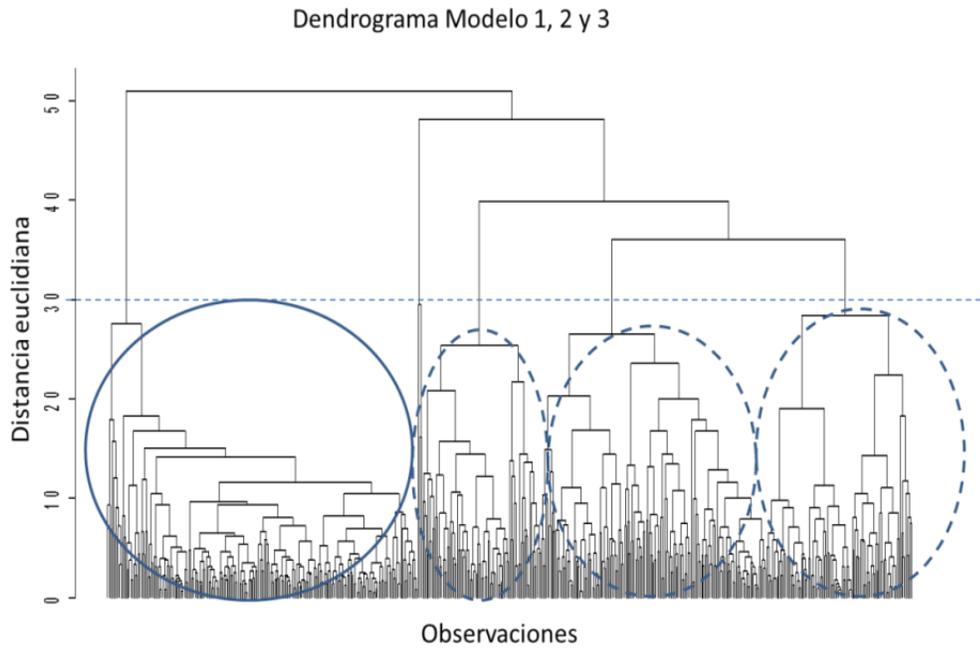


Fig. 4.9 Dendrograma de los modelos uno, dos y tres, para una distancia menor a 40

La Fig. 4.9, muestra el dendrograma de los tres modelos, en donde se observan los grupos que se pueden formar. Si se traza una línea perpendicular al eje de la distancia euclidiana en aproximadamente un valor de 40, el número de grupos formados será tres. Esos tres grupos, pueden representar a cada uno de los modelos y/o cada una de las clases. Si esa línea se desplaza más, se puede observar que se forman otra clasificación con más clases. Se puede observar también la presencia de clases con pocos datos.

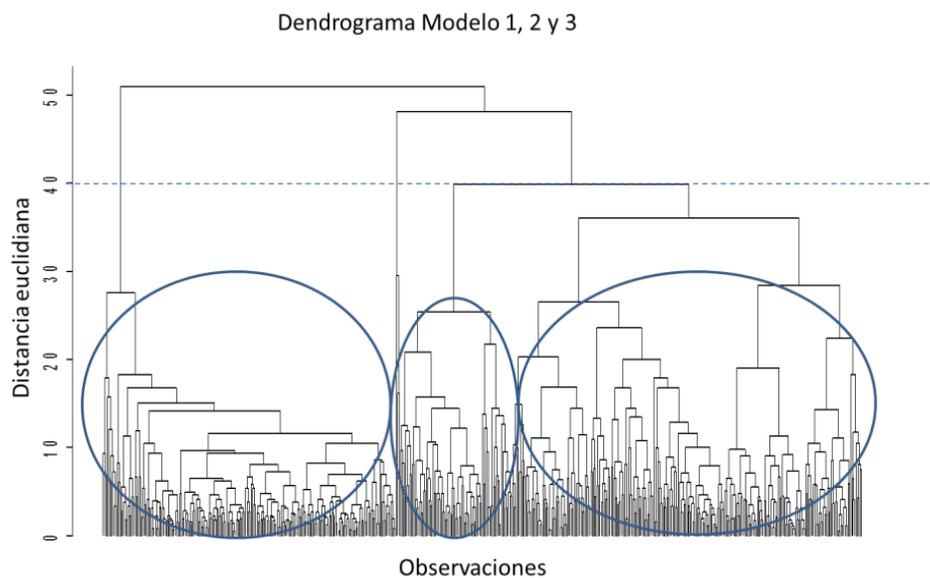


Fig. 4.10 Dendrograma de los modelos uno, dos y tres, para una distancia menor a 40

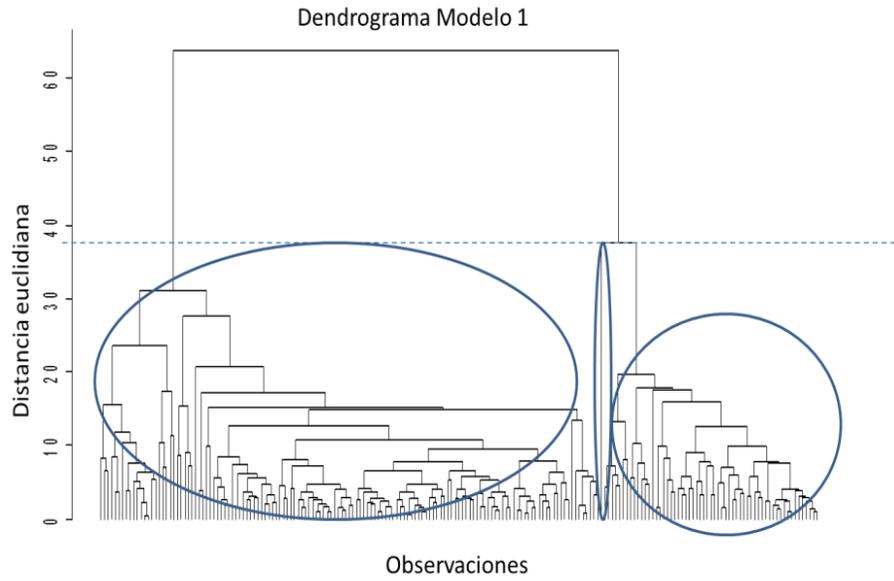


Fig. 4.11 Dendrograma del modelo uno. La línea azul representa una distancia euclidiana de 38

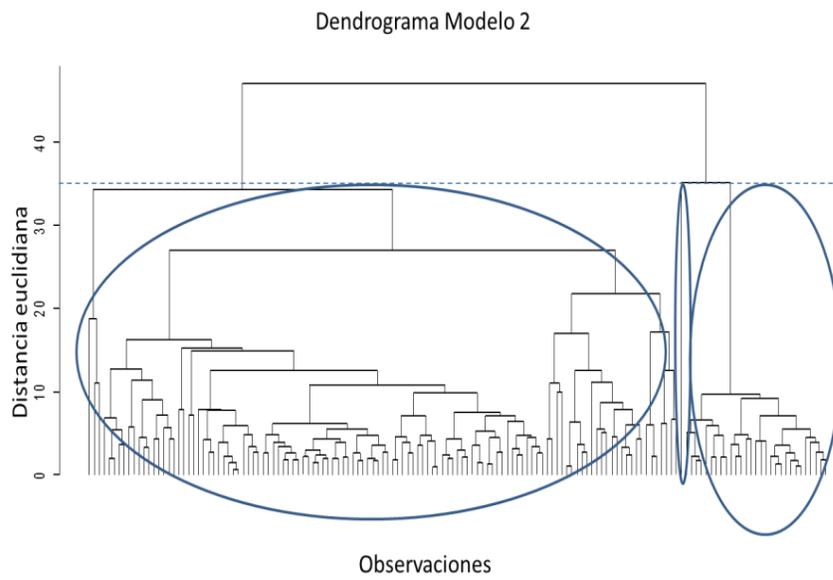


Fig. 4.12. Dendrograma del modelo dos. La línea azul representa una distancia euclidiana de 35

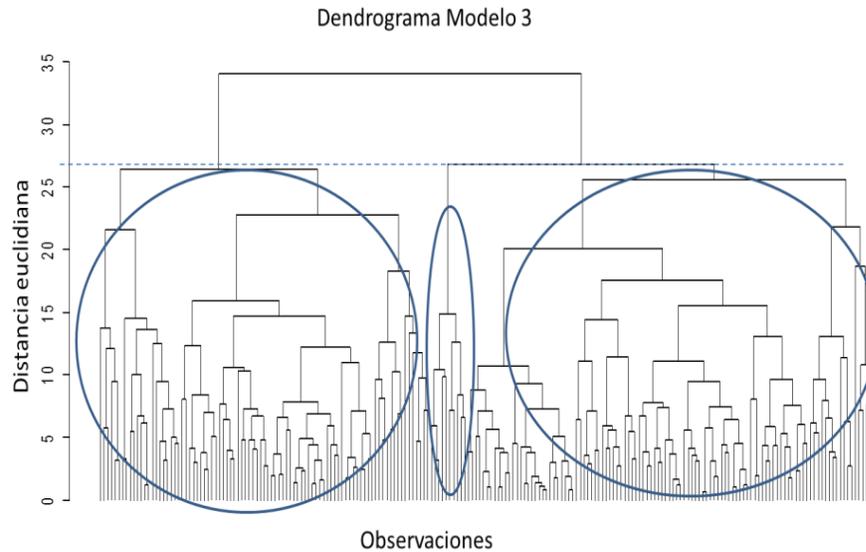


Fig. 4.13. Dendrograma del modelo tres. La línea azul representa una distancia euclidiana de 27

El análisis de dendrogramas se realiza por modelo, los resultados se muestran en las figuras 4.11, 4.12 y 4.13. De este análisis se observa que los grupos formados no son homogéneos, incluso en algunos casos (Modelo uno y 2), es difícil formar los tres grupos (ya que uno tiene muy pocos datos). Lo que permite observar inconsistencia de la clasificación, en función de las variables seleccionadas (todos los registros). Para el siguiente análisis de dendrogramas, se redujo el número de variables, en función del análisis exploratorio previo (Relación RT/GR , Caliper, Neutrón y sónico), mostrando los resultados en las figuras 4.14, 4.15 y 4.16, donde se observa una mejor distribución de los grupos.

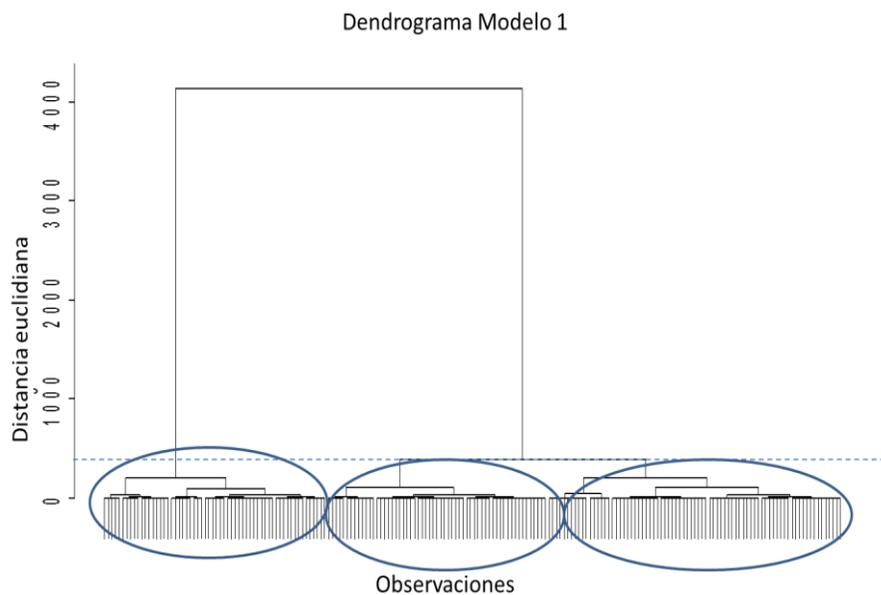


Fig. 4.14. Dendrograma del modelo uno. Utilizando los registros la relación RT/GR , Caliper, Neutrón y sónico

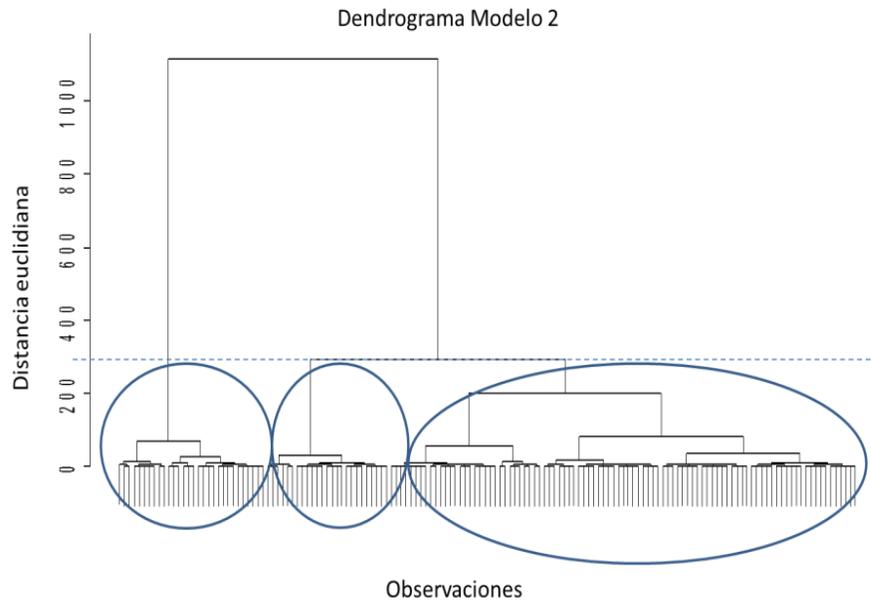


Fig. 4.15. Dendrograma del modelo dos. Utilizando los registros la relación RT/GR , Caliper, Neutrón y sónico

a) K-medias

Apoyados de los dendrogramas antes analizados, se procede a utilizar el algoritmo de las k medias para la clasificación, suponiendo tres clases y los registros relación RT/GR , caliper, neutrón y sónico. Los resultados de dicha clasificación se muestran en las figuras 4.17.

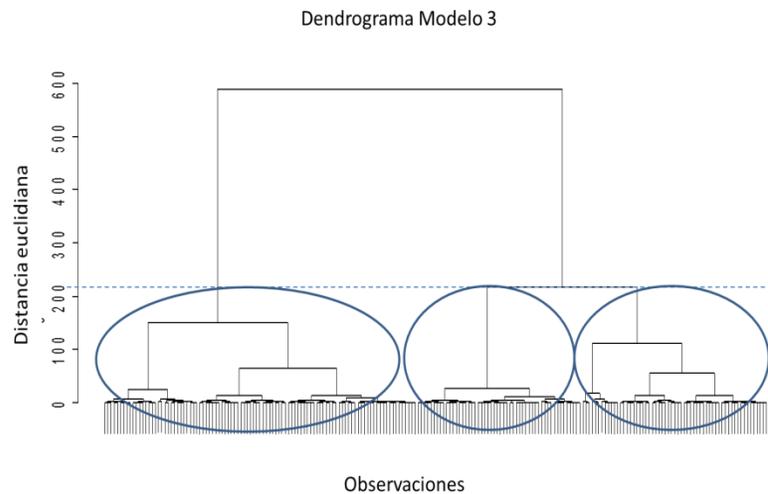


Fig. 4.16. Dendrograma del modelo tres. Utilizando los registros la relación RT/GR , Caliper, Neutrón y sónico

En la Fig. 4.17 se muestran el registro de rayos gama de cada modelo, en el carril uno, así como la clasificación original, en el carril dos, la clasificación por el algoritmo de k medias (carril tres) y en carril cuatro, la clasificación de los registros por k-medias interpretadas. De estos resultados puede observarse que la clasificación de las arenas de la clase uno fue perfectamente reconocida, prácticamente en los tres modelos, en contraste con las facies arcillosas, las cuales presentan mayor problema al ser clasificadas por el algoritmo, y es necesario interpretar los datos para aproximarse a la clasificación original. La

clasificación de k medias interpretada para el modelo tres, es paradójicamente la que más se aproximó a la clasificación real.

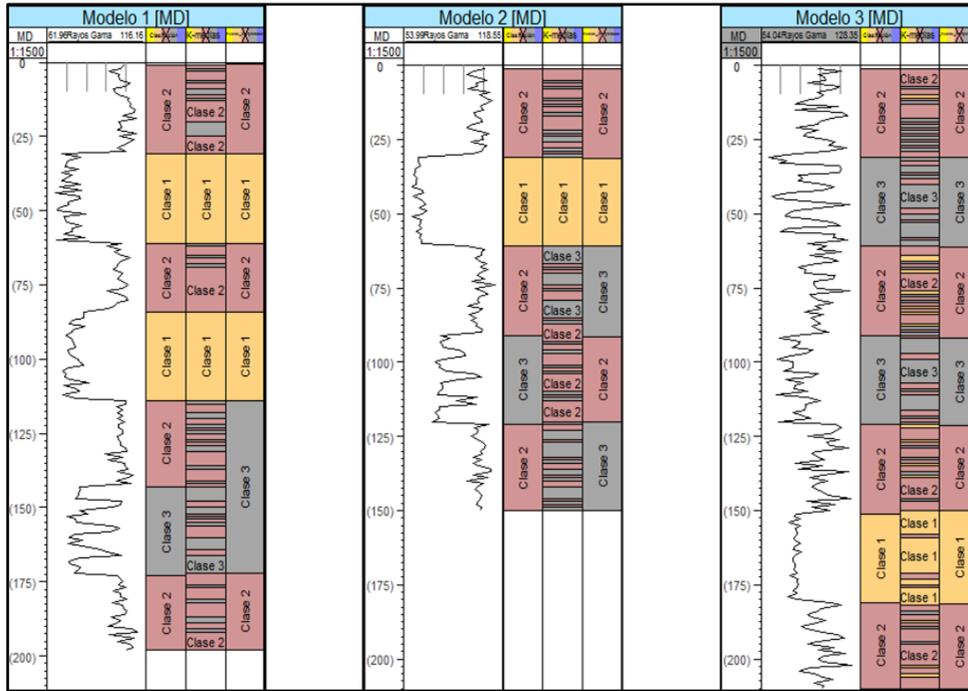


Fig. 4.17. Resultado de la clasificación por el método de k medias.

Las Figs. 4.18, 4.19 y 4.20, muestra gráficos de puntos, para el modelo uno, dos y tres, donde se observan el número de observaciones contra el valor de la variables elegidas (relación RT/GR , sónico, neutrón y, con el objeto de analizar los grupos o clases formadas por el método de k-medias.

Se puede observar que la clase uno es la mejor definida. Los registros sónico, neutrón y la relación RT/GR , son los que mejor definen las clases.

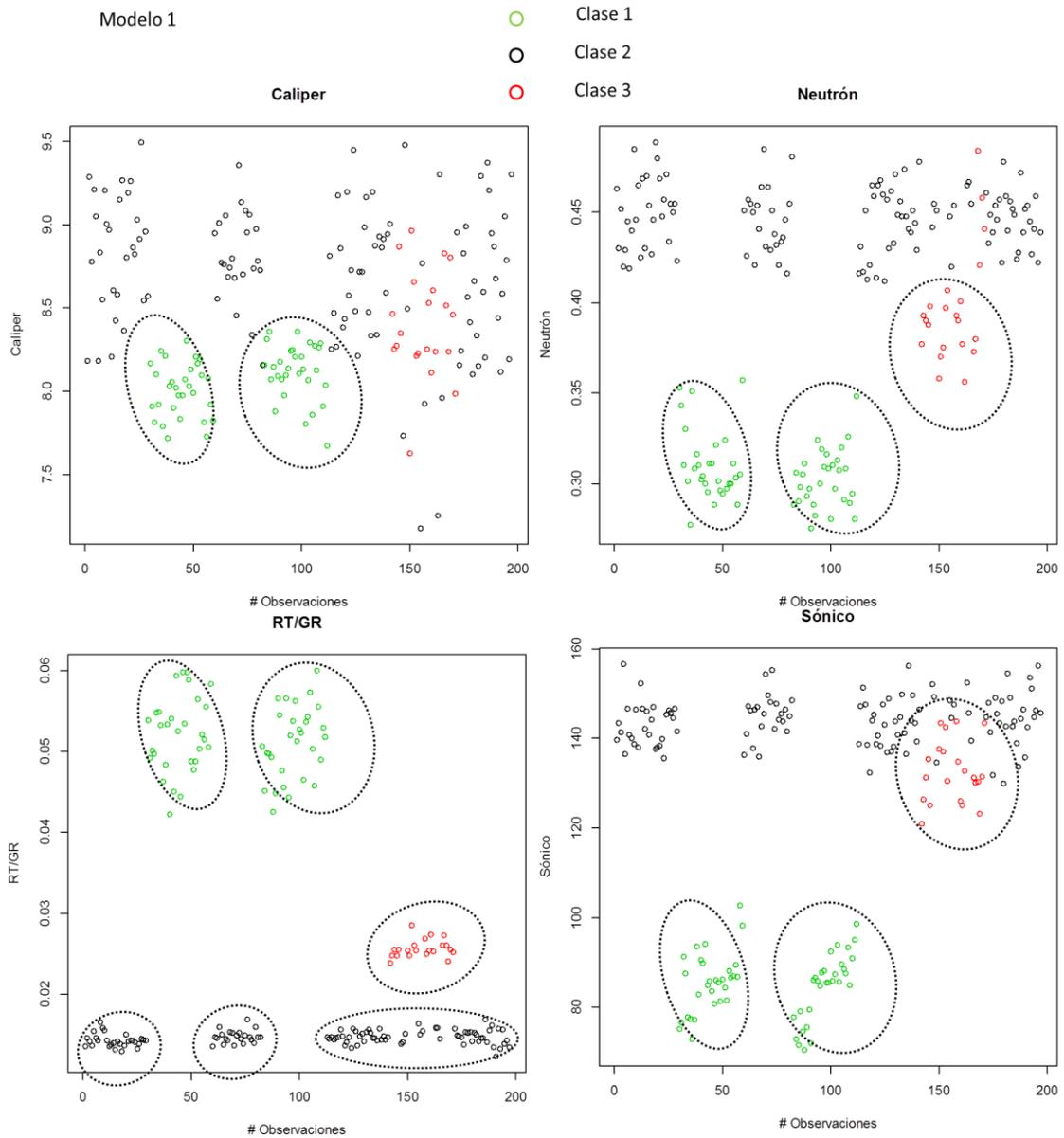


Fig. 4.18. Gráfico de puntos modelo uno, resultado de la clasificación por el método de k medias.

En los gráficos antes citados, se observa que las facies arcillosas son las que más mala clasificación tienen. El modelo tres tiene la más mala clasificación de todas; sin embargo, la parte interpretativa es vital, para definir las clases correctamente.

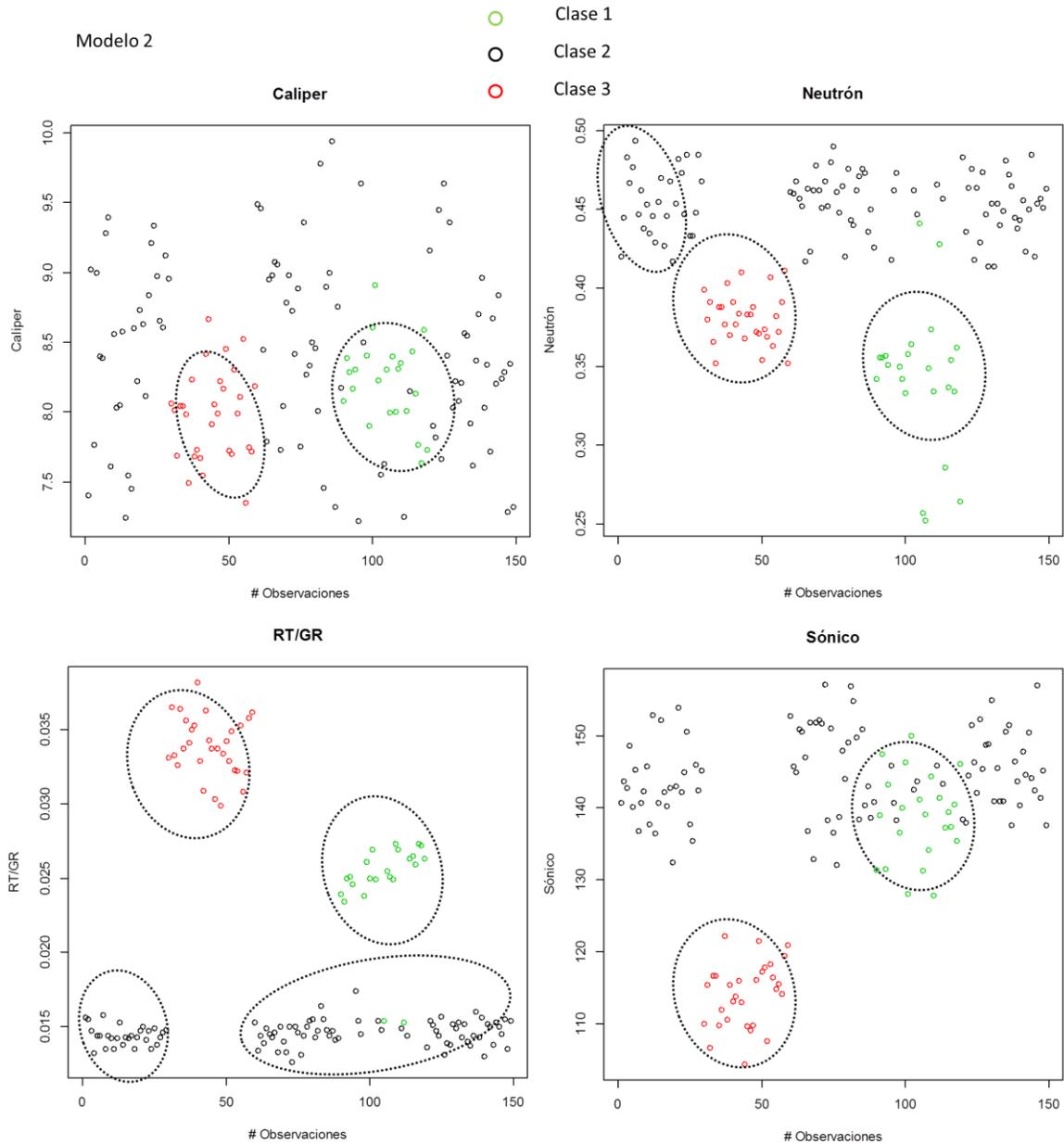


Fig. 4.19. Gráfico de puntos modelo dos, resultado de la clasificación por el método de k medias.

Otros gráficos de apoyo, son los gráficos de dispersión o scatterplot de estas tres variables (Figs. 4.21, 4.22 y 4.23), donde podemos observar los centroides de cada grupo y el centroide total. Los centroides representan el valor medio bivariado y los datos más cercanos a ellos serán clasificados como parte de ese conjunto.

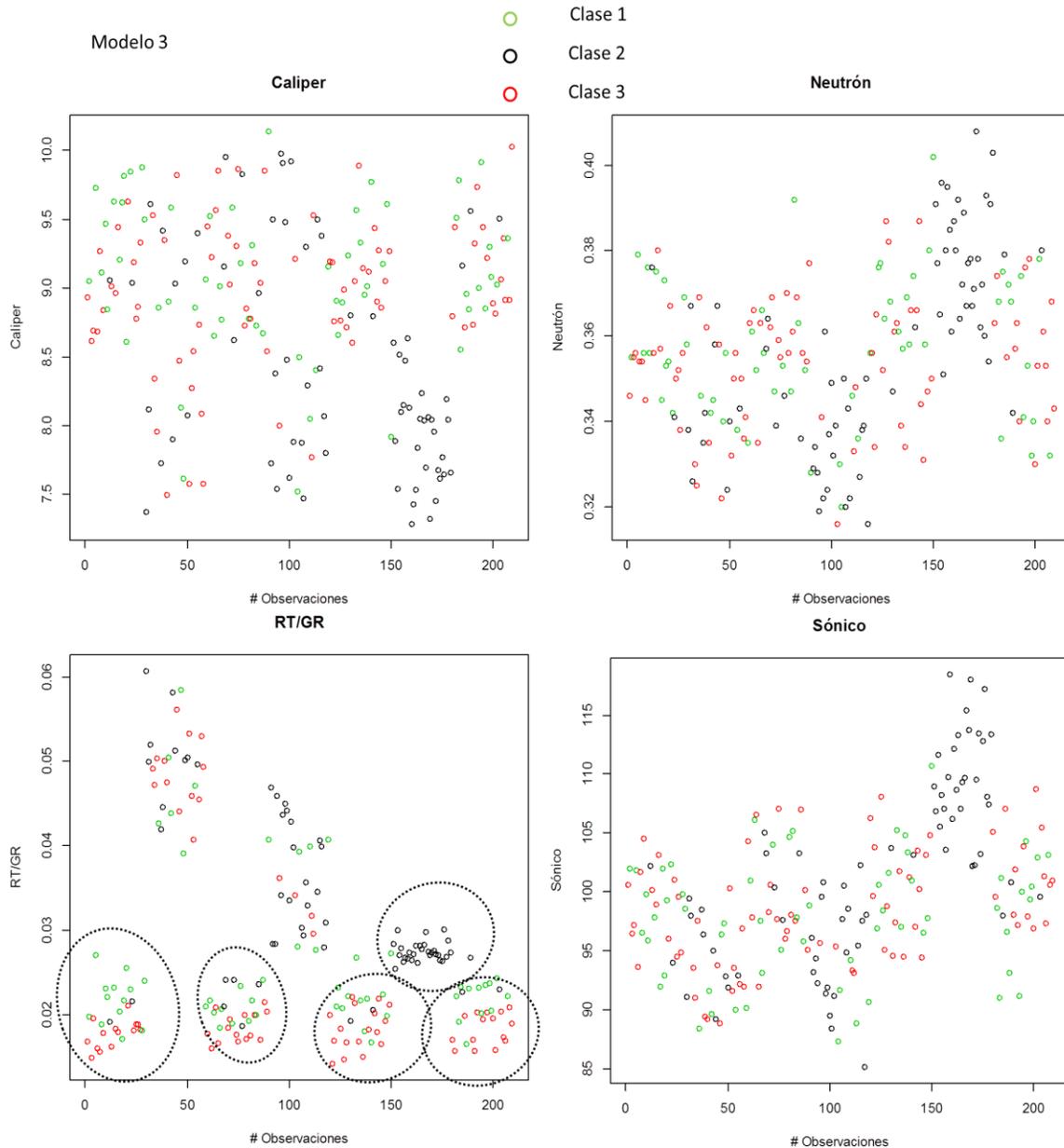


Fig. 4.20. Gráfico de puntos modelo tres, resultado de la clasificación por el método de k medias.

4.4.3.1 Conclusión del Análisis de agrupamiento

Del Análisis de agrupamiento, se puede mencionar que la cantidad de variables, utilizadas en la clasificación, es un factor fundamental, durante los procesos de clasificación. De este análisis junto con el análisis exploratorio de datos pudimos determinar que las variables que mejor clasifican los datos, son la relación RT/GR , el neutrón y sónico. Por lo que se propone un análisis de componentes principales, para establecer si son las variables más óptimas para el entrenamiento de un modelo neuronal.

La clasificación por medio del algoritmo de las k-medias, discrimina perfectamente la clase uno, es decir las arenas limpias, pero tiene problemas al diferenciar las facies arcillosas. Esto debido a que los

registros (modelados) del sónico y neutrón tienen un relativo bajo contraste entre ellas y son más sensitivos al ruido. Se puede calificar los resultados de la clasificación de regular a mala.

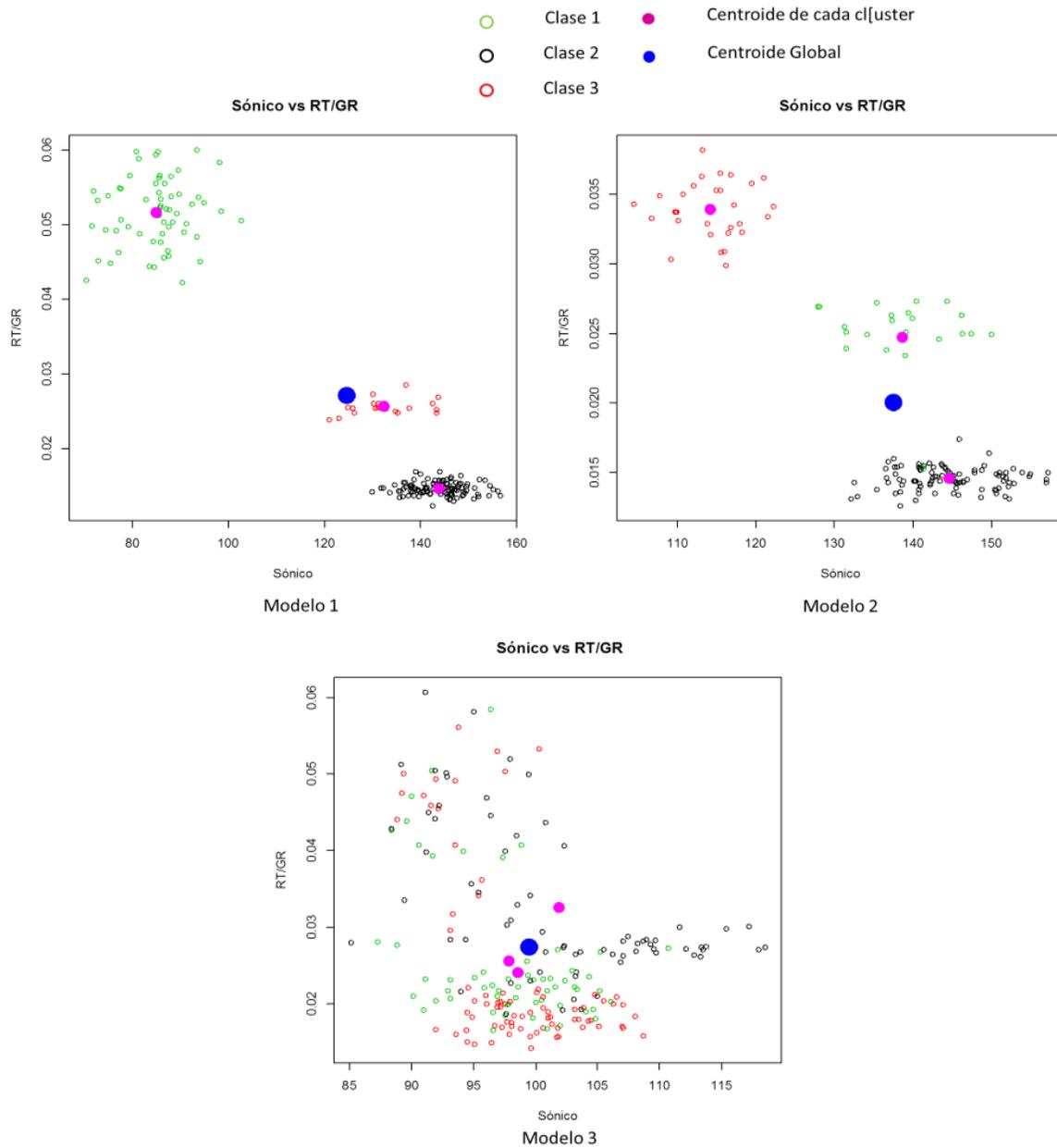


Fig. 4.21. Gráficos de dispersión, para la relación RT/GR y el registro sónico.

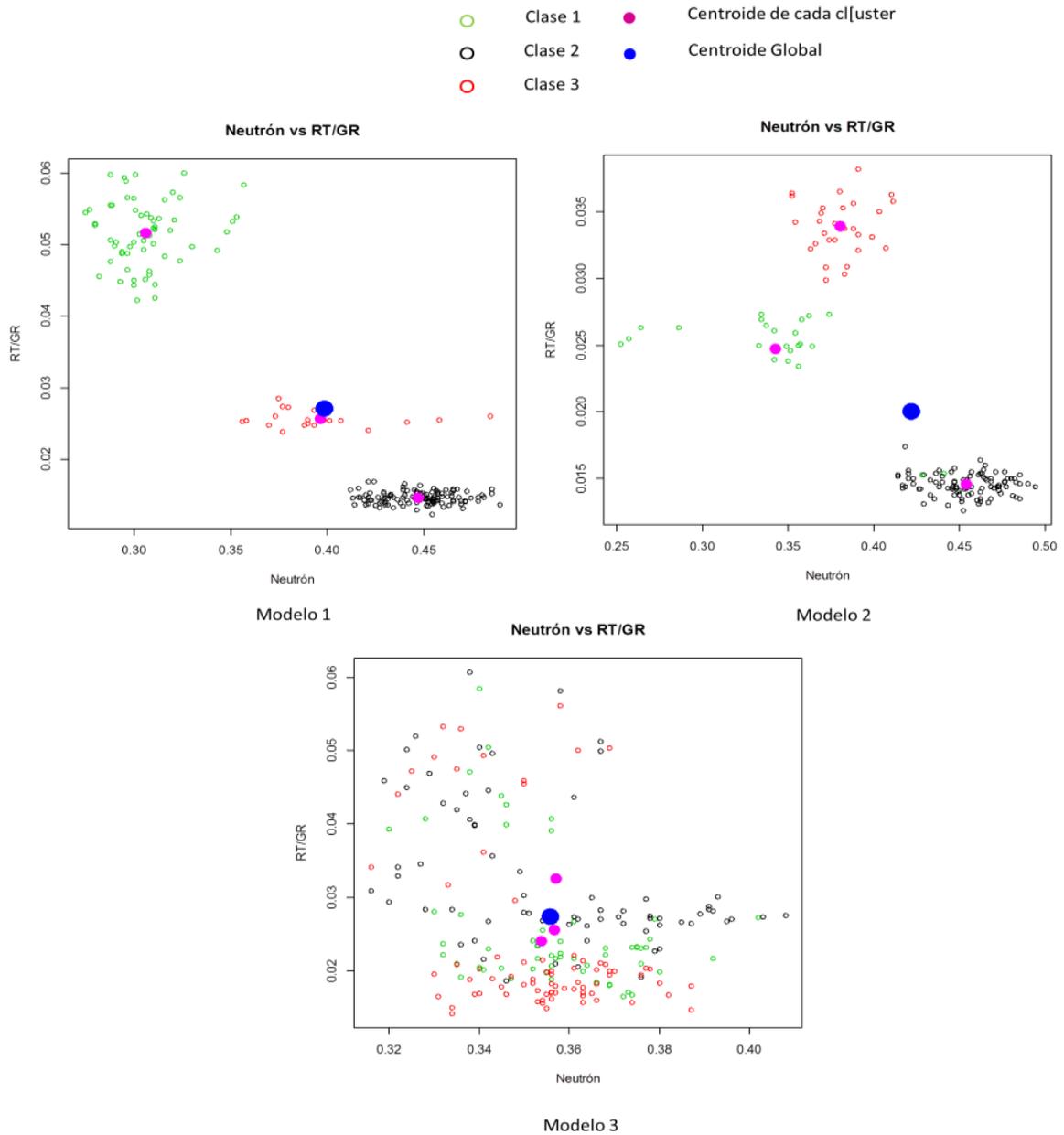


Fig. 4.22. Gráficos de dispersión, para la relación RT/GR y el registro neutrón.

4.4.4 Análisis de componentes principales y selección de variables.

El análisis de componentes principales, es aplicable cuando al analizar los coeficientes de correlación de Pearson, tienden a aproximarse a 1 ó -1, lo que significa mayor grado de correlación lineal. El objetivo como se mencionó en el capítulo 2, es obtener nuevas variables (componentes), que mejor representen las variables originales. Para los registros sintéticos, el análisis de componentes principales se realizó, con todos los datos de todos los modelos, en conjunto y un análisis para cada modelo de forma individual.

Observando los gráficos de dispersión (scatterplots) de los análisis del apéndice A, se puede observar que para el caso del análisis de todos los modelos juntos (uno, dos y tres), se tienen valores de

correlación lineal relativamente bajos, los valores más altos se asocian a los pares *RT/GR*-sónico, *RT/GR*-neutrón y sónico-neutrón.

a) Modelos uno, dos y tres

La Tabla 4.1, muestra las variables que saturan o están más relacionadas con cada una de las componentes, siendo la relación *RT/GR*, la que mayor relación tiene con la componente uno. La Tabla 4.2 muestra los valores propios para cada componente, siendo hasta la componente tres los valores que llegan a 1, y según la regla de Kaiser, serían suficientes tres componentes para representar todas las variables. Lo anterior se muestra también en los gráficos de sedimentación de la Fig. 4.23.

La Tabla 4.3, muestra la proporción de varianza que explica cada componente, por ejemplo la componente uno, representa el 48.3% de la varianza total, la componente dos el 16.9% y así sucesivamente.

Tabla 4.1. Relación de las variables con cada una de las componentes, para los modelos uno, dos y tres

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Caliper	-0.15568	0.61759	0.23311	-0.59116	0.02608	0.43056	0.06571	0.0127
dGR/	-0.084	0.26428	-0.82594	0.1552	-0.37026	0.27909	-0.04096	0.01342
GR	-0.36894	0.37805	-0.25958	0.05396	0.55157	-0.50451	0.11975	-0.2785
Neutron	-0.43937	-0.25981	-0.03631	0.03751	0.36411	0.36642	-0.68467	0.03345
Rho	0.15728	-0.43328	-0.41456	-0.75449	0.14702	-0.14937	0.04357	0.02424
RT	0.44764	0.15498	-0.1444	0.17996	0.56479	0.19686	0.10621	0.59443
RT/GR	0.48614	-0.01554	-0.04325	0.08844	0.25166	0.35193	-0.01351	-0.75257
Sónico	-0.42202	-0.35543	-0.01433	0.11254	0.14656	0.40374	0.70536	-0.02607

Tabla 4.2. Valores propios para cada componente, para los modelos uno, dos y tres

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
3.87115289	1.34927589	1.14741468	0.76963785	0.37927034	0.3202159	0.14021408	0.02281836

Tabla 4.3. Relación de las variables con cada una de las componentes, para los modelos 1

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Desviación estándar	1.96752	1.16158	1.07117	0.87729	0.61585	0.56588	0.37445	0.15106
Proporción de varianza	0.48389	0.16866	0.14343	0.0962	0.04741	0.04003	0.01753	0.00285
Proporción acumulada	0.48389	0.65255	0.79598	0.89219	0.93959	0.97962	0.99715	1

b) Modelo uno

Tabla 4.4. Relación de las variables con cada una de las componentes, para el modelo uno

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Caliper	0.30655	0.19842	-0.0727	0.91746	-0.13192	0.01458	-0.04496	-0.00072
dGR/	0.07485	-0.91682	0.26732	0.15969	-0.23826	0.00577	-0.00858	0.00254
GR	0.401	-0.14801	0.13616	0.03328	0.86759	-0.08229	-0.02932	-0.19311
Neutron	0.41498	0.01562	-0.07231	-0.178	-0.10963	0.87702	-0.0095	-0.09675
Rho	-0.20483	-0.31271	-0.89617	0.09145	0.19789	0.07797	-0.01781	0.05676
RT	-0.41304	0.00353	0.23963	0.16176	0.26775	0.32887	-0.57394	0.48636
RT/GR	-0.42739	0.00973	0.07156	0.13079	0	0.13883	-0.25935	-0.84172
Sónico	0.41498	0.00537	-0.18392	-0.21985	-0.22414	-0.30051	-0.77458	-0.07135

Tabla 4.5. Valores propios para cada componente, para el modelos 1

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
5.27945684	1.05788149	0.82697561	0.54081599	0.14663829	0.08727091	0.03555205	0.02540881

Tabla 4.6. Relación de las variables con cada una de las componentes, para los modelos 1

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Desviación estándar	2.29771	1.02853	0.90938	0.7354	0.38293	0.29542	0.18855	0.1594
Proporción de varianza	0.65993	0.13224	0.10337	0.0676	0.01833	0.01091	0.00444	0.00318
Proporción acumulada	0.65993	0.79217	0.89554	0.96314	0.98147	0.99238	0.99682	1

Las Tablas 4.4, 4.5 y 4.6, muestran los resultados del análisis de componentes principales para el modelo uno, donde se observa un comportamiento similar al análisis global de todos los modelos en, donde se puede ver que la relación RT/GR es también el que más peso tiene en la componente uno y que se requerirían 2 componentes para explicar todas las demás variables. La componente uno explica el 66% de la varianza total.

c) Modelo dos

Tabla 4.7. Relación de las variables con cada una de las componentes, para el modelo dos

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Caliper	-0.19374	0.11242	0.86847	-0.43353	-0.08664	0.01104	0.00308	-0.00009
dGR/	-0.11953	-0.66009	0.37087	0.63536	-0.05161	0.01386	-0.06757	0.03669
GR	-0.45394	-0.01261	-0.08873	0.04197	-0.04707	0.25124	0.59176	-0.60715
Neutron	-0.37721	-0.30689	-0.10945	-0.2629	0.59287	-0.57004	0.05319	0.05621
Rho	0.23155	-0.6138	-0.17514	-0.51441	-0.49726	-0.05494	0.15507	-0.00338
RT	0.42525	0.16485	0.18233	0.21706	-0.0291	-0.45562	0.69215	0.15532
RT/GR	0.45484	-0.03436	0.12738	0.0055	0.14415	-0.26368	-0.30099	-0.77122
Sónico	-0.39813	0.22846	-0.09068	0.1604	-0.60599	-0.57574	-0.2207	-0.0891

Tabla 4.8. Valores propios para cada componente, para el modelo dos

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
4.59145849	1.14659854	0.88190131	0.77320117	0.35471567	0.15306618	0.08426191	0.01479674

Tabla 4.9. Relación de las variables con cada una de las componentes, para el modelo dos

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Desviación estándar	2.14277	1.07079	0.9391	0.87932	0.59558	0.39124	0.29028	0.12164
Proporción de varianza	0.57393	0.14332	0.11024	0.09665	0.04434	0.01913	0.01053	0.00185
Proporción acumulada	0.57393	0.71726	0.82749	0.92414	0.96848	0.98762	0.99815	1

Las Tablas 4.7, 4.8 y 4.9, muestran los resultados del análisis de componentes principales para el modelo dos, donde se observa un comportamiento similar al análisis de todos los modelos en conjunto, donde nuevamente la relación RT/GR el que más peso tiene la componente uno y que al igual que el modelo uno se requerirían dos componentes para explicar todas las demás variables. La componente uno explica el 87.3% de la varianza total.

c) Modelo tres

Tabla 4.10. Relación de las variables con cada una de las componentes, para el modelo tres

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Caliper	-0.19524	-0.43227	0.2796	-0.61046	-0.1254	0.54841	-0.08669	-0.01409
dGR/	-0.06946	-0.31524	-0.6537	0.22176	0.06477	0.22574	-0.60281	-0.02745
GR	-0.22698	-0.513	-0.38397	-0.0047	-0.1827	-0.23537	0.58656	0.32408
Neutron	-0.34271	0.39198	-0.13733	-0.10716	-0.81422	-0.07537	-0.17237	0.0144
Rho	0.07584	0.3169	-0.46102	-0.72887	0.31398	-0.22674	0.00621	-0.00094
RT	0.55596	-0.07643	-0.2483	-0.01719	-0.32193	0.20449	0.30898	-0.61822
RT/GR	0.59708	0.12267	-0.04958	-0.02204	-0.18687	0.28011	-0.03502	0.71477
Sónico	-0.34167	0.41908	-0.22706	0.1861	0.21586	0.64461	0.39826	0.0268

Tabla 4.11. Valores propios para cada componente, para el modelo tres

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
2.5407092	1.8586816	1.4173641	0.7888196	0.5531977	0.4642889	0.3648109	0.012128

Tabla 4.12. Relación de las variables con cada una de las componentes, para el modelo tres

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Desviación estándar	1.59396	1.36333	1.19053	0.88816	0.74377	0.68139	0.604	0.11013
Proporción de varianza	0.31759	0.23234	0.17717	0.0986	0.06915	0.05804	0.0456	0.00152
Proporción acumulada	0.31759	0.54992	0.72709	0.8257	0.89485	0.95288	0.99848	1

Las Tablas 4.10, 4.11 y 4.12, muestran los resultados del análisis de componentes principales para el modelo tres, donde se observa un comportamiento similar al análisis de todos los demás modelos. Para

este caso se requerirían tres componentes para explicar todas las demás variables. La componente uno explica el 31.8 % de la varianza total.

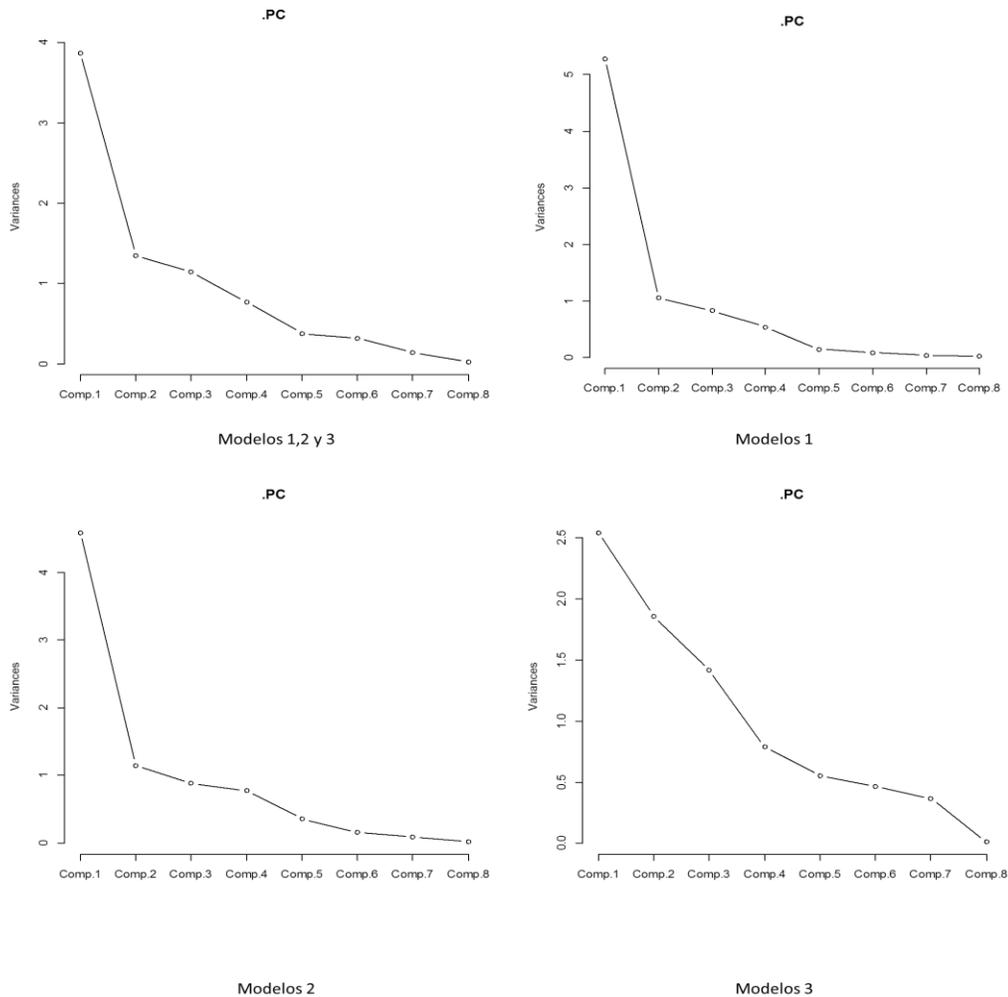


Fig. 4.23. Gráficos de sedimentación, para los análisis de modelos en conjunto, y para los modelos uno, dos y tres de manera individual.

d) Análisis de componentes principales y análisis de grupos (clústeres)

Mediante el algoritmo de componentes principales y el algoritmo de k-medias, se generaron biplots para cada modelo de manera individual. Los biplots muestran la relación que guardan las dos primeras componentes principales y las variables utilizadas en el análisis. Para este análisis, se realizó el Análisis de agrupamiento por k-medias, definiendo tres grupos y graficando los biplots, haciéndolo en dos etapas:

1. Utilizando todos los registros rayos gama, resistividad, caliper densidad, sónico, neutrón, la relación RT/GR y la derivadas de los rayos gama con respecto de la profundidad (biplots del lado izquierdo de la Fig. 4.24 A)
2. Utilizando únicamente los registros caliper, sónico, neutrón y la relación RT/GR (biplots del lado derecho de la Fig. 4.24 B)

La Fig. 4.24, muestra del lado izquierdo los biplots asociados a cada modelo utilizando todos los registros. Los puntos etiquetados por su clase determinada por el análisis de k medias (clase uno, dos y tres), se interpretan en función de la distancia. Los puntos más cercanos representan puntos con relaciones similares para las componentes uno y dos, así como valores similares de las variables utilizadas para cada análisis (representadas como vectores). Los puntos que caen cerca del origen de los vectores, se aproxima a la media de cada variable. Los puntos más alejados de los vectores indican valores bajos, los puntos cercanos a los vectores, valores más altos, por lo que se pueden detectar valores atípicos (outliers). A partir de la interpretación de los puntos se pueden analizar grupos. La magnitud y dirección de los vectores, están relacionados con el grado de correlación multivariada entre las variables. Propiamente dicho la longitud está relacionada con la variabilidad y los ángulos que forman dos vectores con la covariabilidad (correlación). La dirección está relacionada con el grado de correlación de los componentes principales y principalmente marca tendencias.

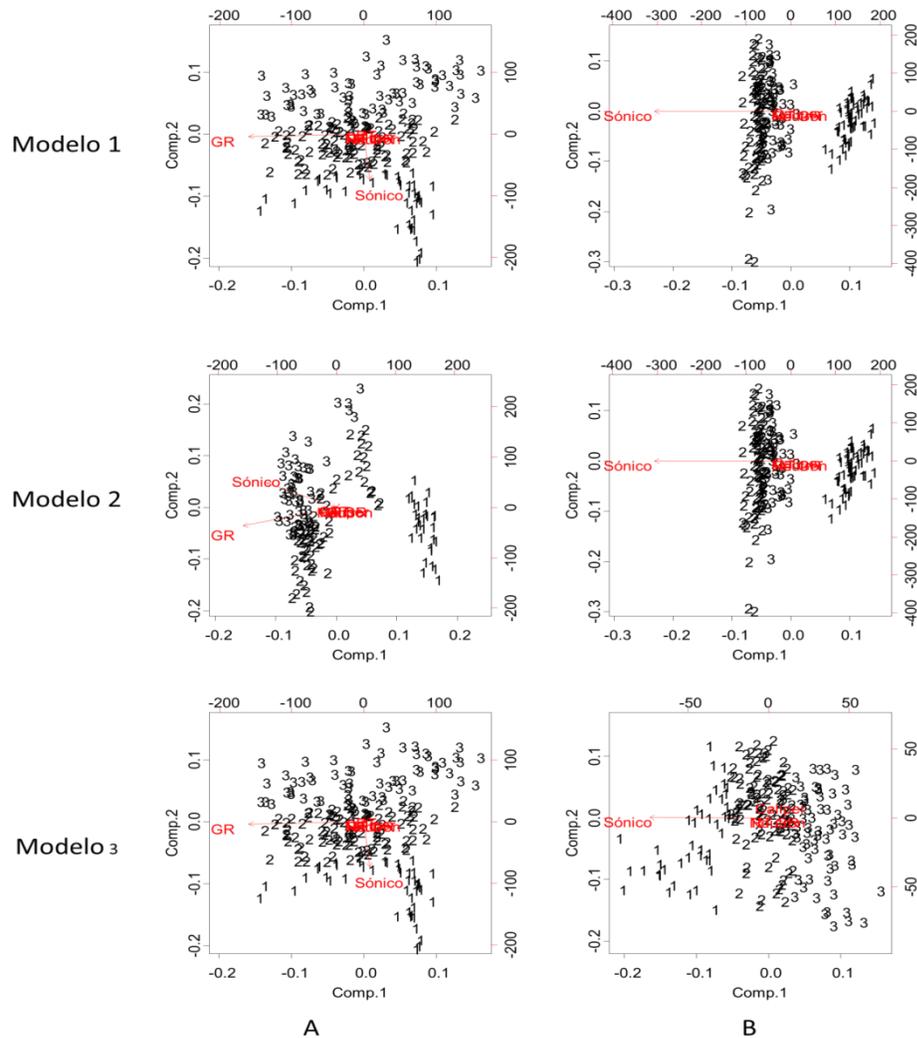


Fig. 4.24. Biplots (a la izquierda se muestran los gráficos realizando el análisis de k-medias con todos los registros y a la derecha utilizando solo neutrón, sónico, RT/GR y caliper)

En la Fig. 4.24 se muestra el grado de correlación entre el sónico y rayos gama, siendo estos los de más alta variabilidad. Los demás presentan bajas varianzas, debido a que la longitud del vector es muy

pequeña. Para los modelos uno y tres las correlaciones, entre el registro de rayos gama y el sónico, son fuertes, mientras que para el modelo dos, las correlaciones son más bajas. Se puede apreciar nuevamente que la clase uno, es la que más fácilmente puede identificarse (a excepción del modelo tres) usando los registros de rayos gama, caliper, densidad y sónico, que tienden a identificar las facies arcillosas (debido a que responden al grado de compactación, presión de poro, contenido mineralógico, etc.), mientras que la resistividad y neutrón, tienden a identificar facies más arenosas (contenido de fluidos) mientras que los demás registros tienden a responder a las tres clases.

4.4.4.1 Conclusiones del análisis de componentes principales.

Como conclusiones de este análisis, se tiene que los registros que mejor definieron los grupos al realizar la clasificación por medio el algoritmo de k-medias, son los que muestran mayor grado de correlación, por lo que influyen menos en las componentes principales. En teoría las componentes que deberían clasificar mejor los datos son las que utilizan la mayor cantidad de información y representan la mayor variabilidad de los datos, sin embargo para definir los grupos se observó que el análisis de componentes principales usando solo caliper, *RT/GR*, neutrón y sónico, es la mejor opción; por lo que se proponen tres esquemas para las pruebas de entrenamiento de las redes neuronales:

- 1) Utilizar el análisis de componentes principales usando todos los datos, para entrenar la red
- 2) Utilizar el análisis de componentes principales usando solo los registros caliper, *RT/GR*, neutrón y sónico, sin ningún tipo de transformación.
- 3) Sin utilizar el análisis de componentes principales, únicamente los registros caliper, *RT/GR*, neutrón y sónico, sin ningún tipo de transformación.

4.4.5 Redes neuronales

Se aplicaron los diferentes modelos neuronales perceptron multicapa (MLP), mapas auto-organizados (SOM) y máquinas comité (CM), mediante la programación en lenguaje M de Matlab

4.4.5.1 Redes perceptrón multicapa.

Se generaron diversas pruebas, modificando los parámetros de entrenamiento principalmente el número de capas y neuronas. Se entrenaron tres redes neuronales cada uno de ellos utilizando un modelo en específico y estimando los otros dos.

a) Entrenado con el modelo uno (bien clasificado)

Los parámetros utilizados fueron:

- Dos capas de ocultas con cinco neuronas la primera y una segunda con tres neuronas.
- Las funciones de transferencia fueron "tansig" para la primera capa "logsig" para la segunda y "purelin" para la capa de salida.
- Para definir los datos de entrada, se realizaron diferentes pruebas y apoyado con el análisis exploratorio de datos, se optó por utilizar los registros sónico, neutrón, caliper y la relación *RT/GR*.
- Se normalizaron los datos, entre cero y uno.

- Se realizó una segunda prueba entrenando la red utilizando las dos primeras componentes principales.
- Se dividieron los datos de entrenamiento, validación y prueba, en 60%, 20% y 20%, respectivamente.

Las figuras 4.25, 4.26, 4.27 y 4.28, resumen el proceso de entrenamiento y estimación.

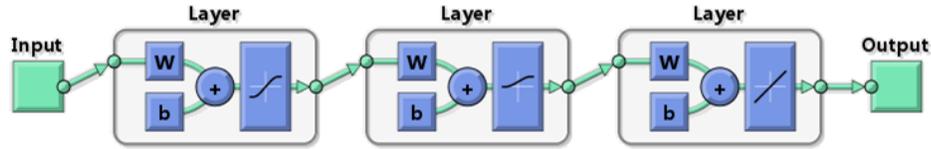


Fig. 4.25. Modelo de red neuronal

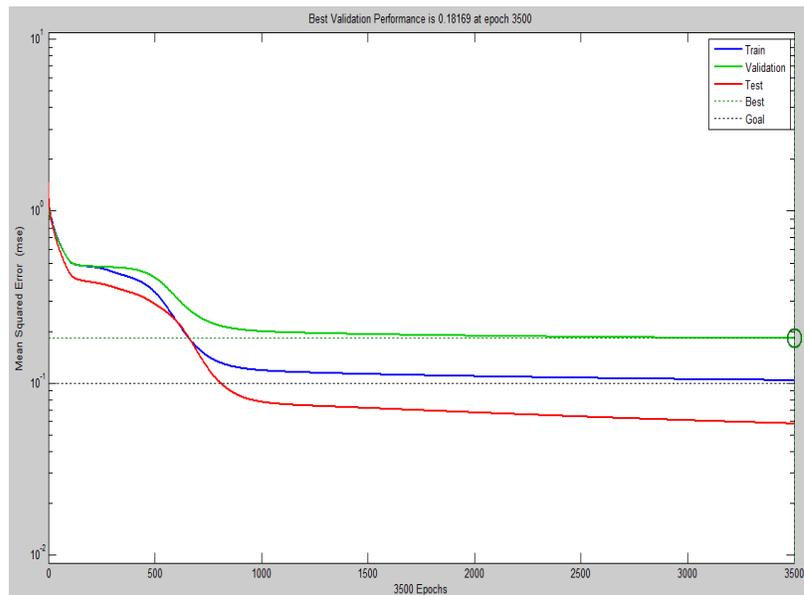


Fig. 4.26. Gráfico del monitoreo del error cuadrático medio, durante cada época (epoch)

En la Fig. 4.28, se muestra la matriz de confusión de la estimación del modelo uno, entrenando con el modelo uno, donde se puede ver que la red predijo prácticamente el 100% de los datos correctamente, a excepción del 4.5 % del total, que representa el 30% de la clase tres. Los resultados de la clasificación se muestran en las Figs. 4.34 A y 4.34 D (utilizando componentes principales). Es importante mencionar, que una reinterpretación manual fue necesaria mostrando los resultados en la Fig. 4.35 A

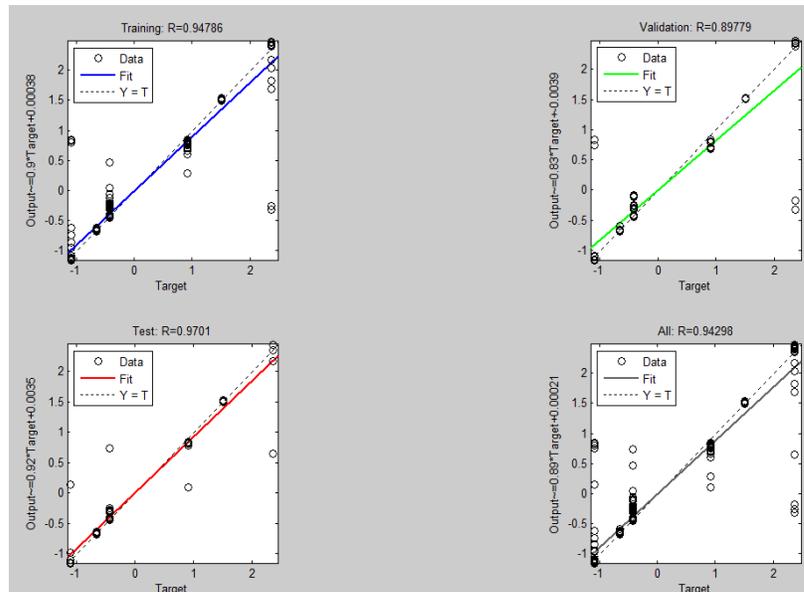


Fig. 4.27. Cuantificación del coeficiente de Pearson y gráfico de regresión lineal de los objetivos y las salidas para los diferentes sets de datos (de izquierda a derecha y de arriba abajo, set de entrenamiento, set de validación, set de prueba y una salida global)

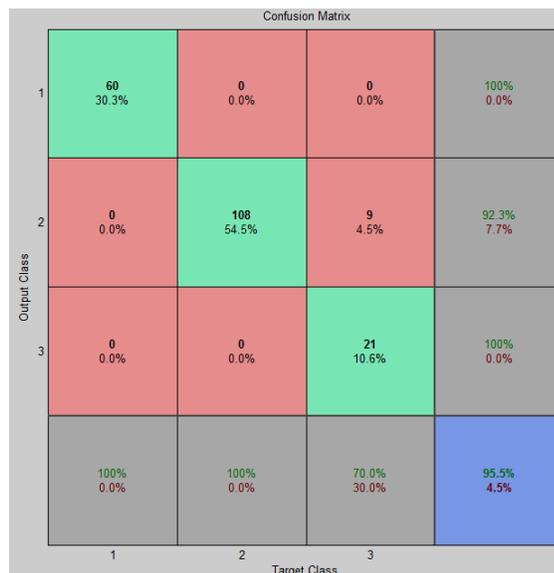


Fig. 4.28. Matriz de confusión, de la estimación del modelo uno, entrenando con el modelo uno

b) Entrenado con el modelo dos (medianamente clasificado)

Los parámetros utilizados fueron:

- Dos capas de ocultas con cinco neuronas la primera y una segunda con tres neuronas.
- Las funciones de transferencia fueron “tansig” para la primera capa, “logsig” para la segunda y “purelin” para la capa de salida.

- Se utilizaron los mismos registros que en la prueba anterior (sónico, neutrón, caliper y la relación RT/GR .)
- Se normalizaron los datos, entre cero y uno.
- Se dividieron los datos de entrenamiento, validación y prueba, en 60%, 20% y 20%, respectivamente.

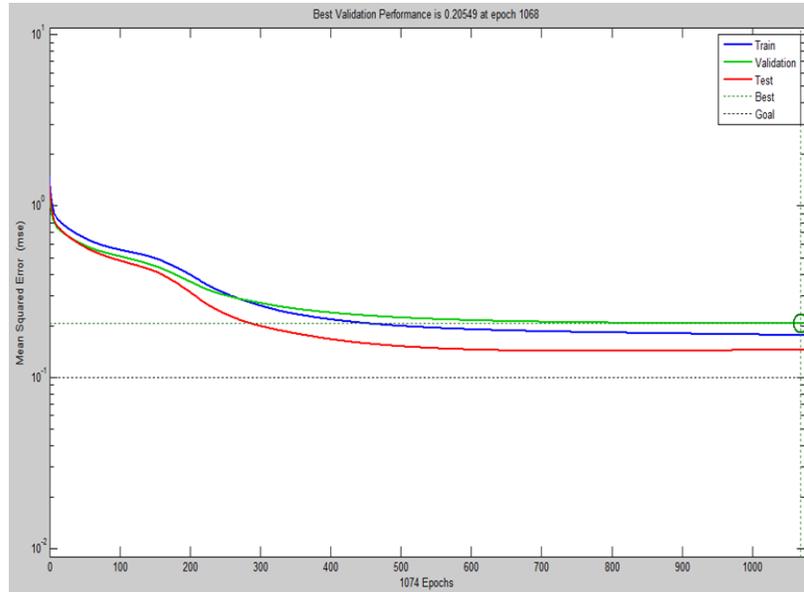


Fig. 4.29. Gráfico del monitoreo del error cuadrático medio, durante cada época (epoch)

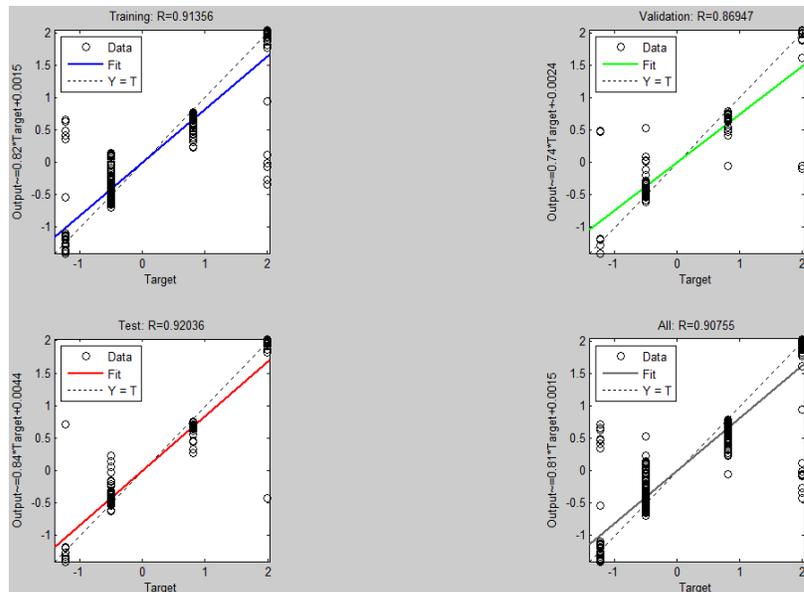


Fig. 4.30. Cuantificación del coeficiente de Pearson y gráfico de regresión lineal de los objetivos y las salidas para los diferentes sets de datos (de izquierda a derecha y de arriba abajo, set de entrenamiento, set de validación, set de prueba y una salida global)

Las figuras 4.29, 4.30 y 4.31, resumen el proceso de entrenamiento y estimación.

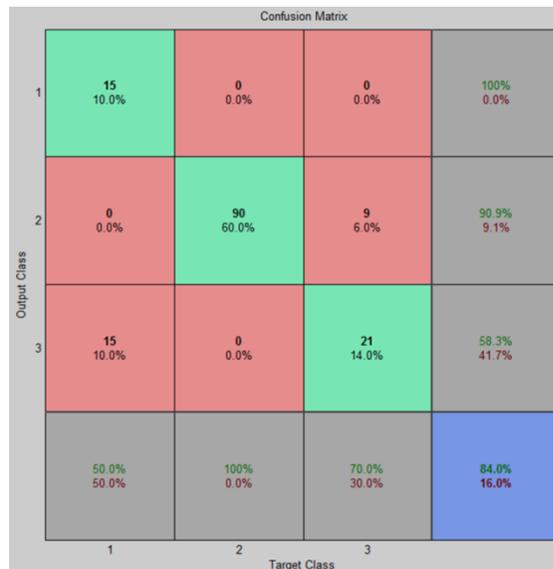


Fig. 4.31. Matriz de confusión, de la estimación del modelo dos, entrenando con el modelo uno

En la Fig. 4.31, se muestra la matriz de confusión de la estimación del modelo dos, entrenando con el modelo uno, donde se puede observar que la red estimó correctamente el 10% como clase uno, el 60% como clase dos y el 14% como clase tres, se puede observar también que estimó el 50% de los datos de la clase uno bien y el otro 50%, lo estimó como clase tres. Estimó correctamente la clase dos y estimó correctamente el 70% de la clase tres, el resto de esta clase (30%) fue estimada como clase dos. Finalmente la red estimó el 84% del total de los datos correctamente, contra un 16% que fue estimado erróneamente. Se observa además que los errores son mayores en los sets de prueba, con un grado de generalización aceptable. Los resultados de la clasificación se muestran en la Fig. 4.35 B y 4.36 B.

c) Entrenado con el modelo tres (mal clasificado)

Los parámetros utilizados fueron:

- Dos capas de ocultas con 5 neuronas la primera y una segunda con 3 neuronas.
- Las funciones de transferencia fueron “tansig” para la primera capa, “logsig” para la segunda y “purelin” para la capa de salida.
- Se utilizaron los mismos registros que en la prueba anterior (sónico, neutrón, caliper y la relación *RT/GR*.)
- Se normalizaron los datos, entre cero y uno.
- Se dividieron los datos de entrenamiento, validación y prueba, en 60%, 20% y 20%, respectivamente.

Las figuras 4.32, 4.33 y 4.34, resumen el proceso de entrenamiento y estimación.

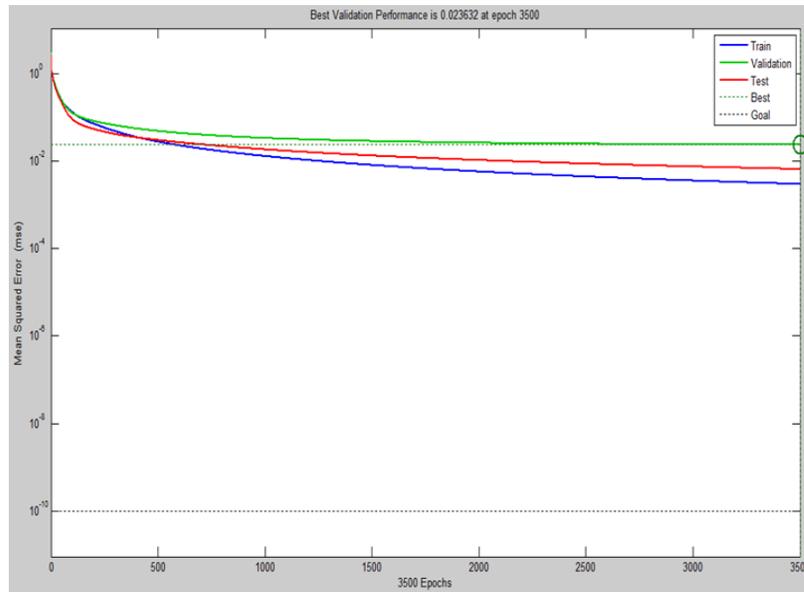


Fig. 4.32. Gráfico del monitoreo del error cuadrático medio, durante cada época (epoch)

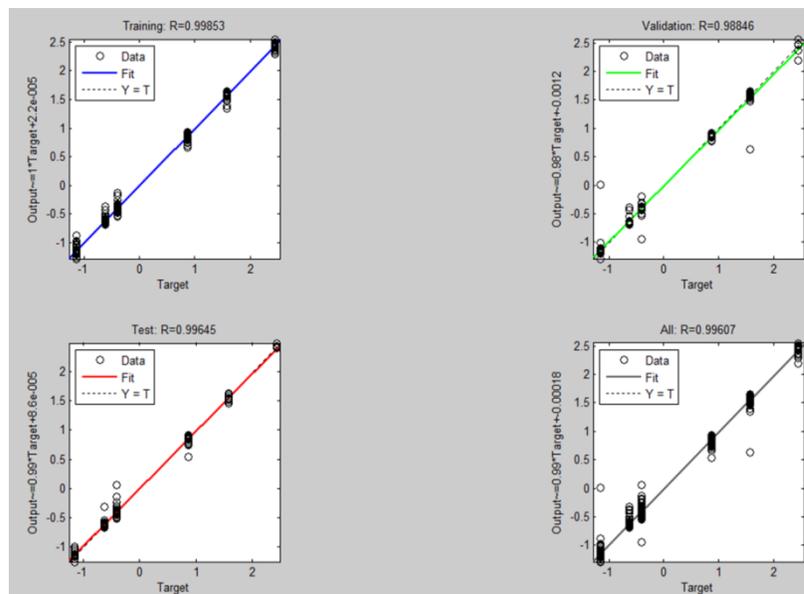


Fig. 4.33. Cuantificación del coeficiente de Pearson y gráfico de regresión lineal de los objetivos y las salidas para los diferentes sets de datos (de izquierda a derecha y de arriba abajo, set de entrenamiento, set de validación, set de prueba y una salida global)

En la Fig. 4.34, se muestra la matriz de confusión de la estimación del modelo tres, entrenando con el modelo uno, donde se puede ver observar que la red estimó correctamente el 1% de los datos como clase uno, el 3.8 % como clase dos y 0% como clase tres. Se observa que se estimó el 6.7% de los datos de la clase uno bien y el 93.3 % restante, fue estimado como clase tres. Se estimó también correctamente el 6.7% de la clase dos y del 93.3% restante, fue estimado como clase uno principalmente y clase dos. La red no pudo estimar correctamente la clase tres, la cual fue estimada como clase uno. Finalmente la red confundió un 95.2% del total de los datos, contra un 4.8% que fue estimado correctamente.

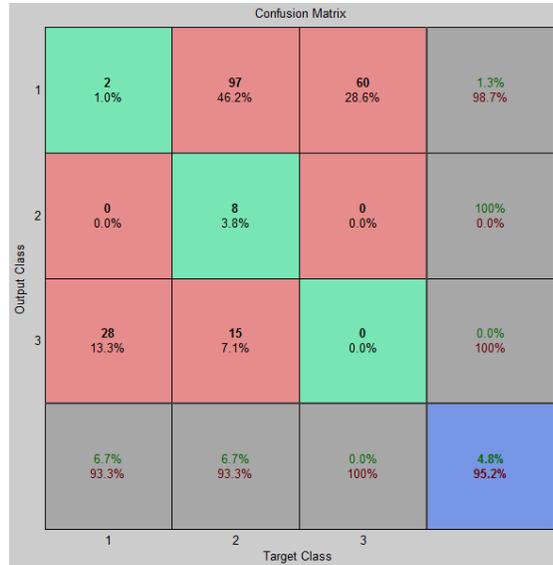


Fig. 4.34. Matriz de confusión, de la estimación del modelo3, entrenando con el modelo1

4.4.5.2 Resultados de la clasificación usando la red Perceptron Multicapa (MPL)

La Fig. 4.35 muestra la clasificación de cada uno de los modelos, entrenando la red con los modelos uno, dos y tres. Fig. 4.36, muestra una reinterpretación de esos resultados, en ella se puede observar que entrenando con el modelo uno (con y sin componentes principales), se puede estimar muy bien el mismo modelo uno y relativamente bien el modelo dos (la red se confunde únicamente en la clase uno). Para el entrenamiento con el modelo dos, se encontró una buena aproximación para estimarse a sí mismo y para estimar el modelo uno. Para el caso del modelo tres, el único resultado satisfactorio es para estimarse así mismo, lo cual se previó, durante el análisis exploratorio de datos.

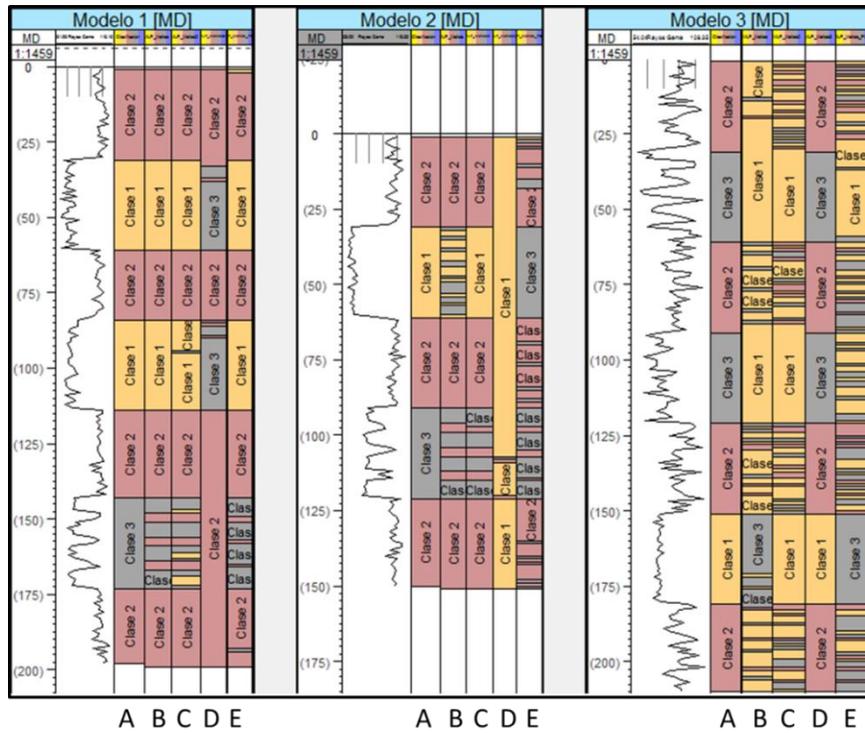


Fig. 4.35. Resultados de la clasificación usando *MPL*. A) Clasificación real, B) Entrenando con modelo uno, C) Entrenando Modelo dos, D) Entrenando Modelo tres, E) Entrenando Modelo uno con componentes principales

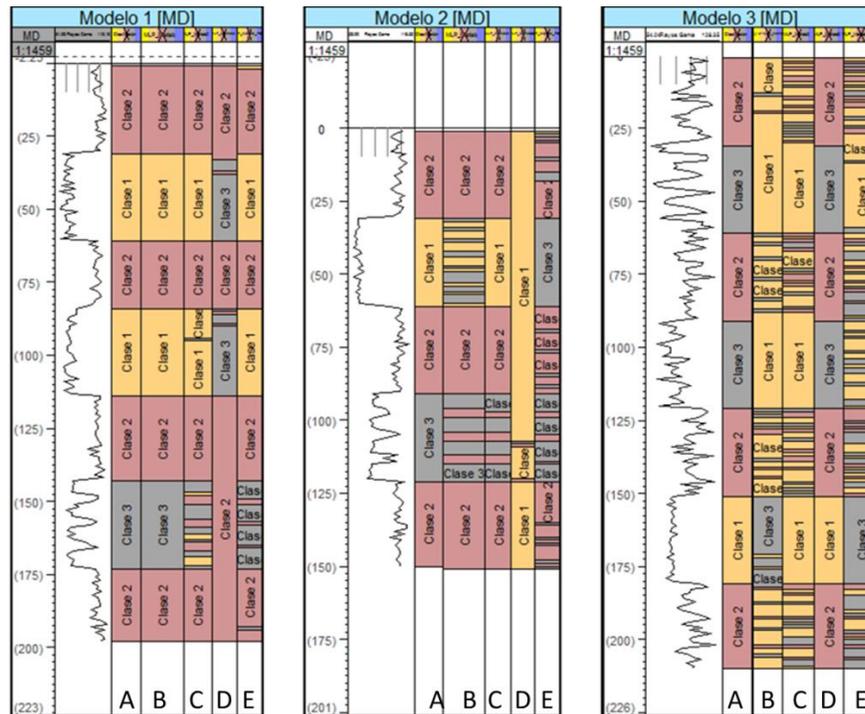


Fig. 4.36. Resultados de la clasificación, reinterpretando las salidas de la red *MPL*. A) Clasificación real B) Entrenando con modelo uno, C) Entrenando Modelo dos, D) Entrenando Modelo tres, E) Entrenando Modelo uno con componentes principales

4.4.5.3 Mapas auto-organizados.

Para el entrenamiento con los mapas auto-organizados, se procedió de igual manera que para las redes *MPL*, se entrenó la red con cada modelo, para posteriormente tratar de estimar los otros dos.

s) Entrenado con el modelo uno (bien clasificado)

Los parámetros utilizados fueron:

- Un mapa de 10 x 10 neuronas (100 neuronas), en una topología hexagonal.
- Se utilizaron únicamente los datos normalizados entre cero y uno.

La Fig. 4.37 muestra los pesos o distancia que tienen los datos para una determinada vecindad de neuronas.

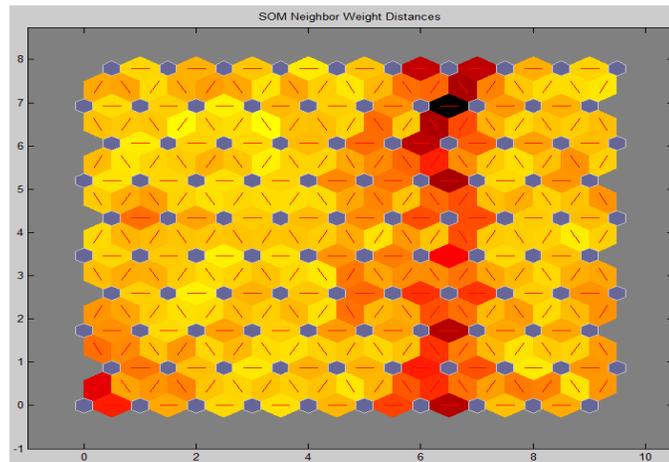


Fig. 4.37. Mapa de pesos para el modelo uno, donde cada color representa el peso que le da una vecindad de neuronas a las entradas

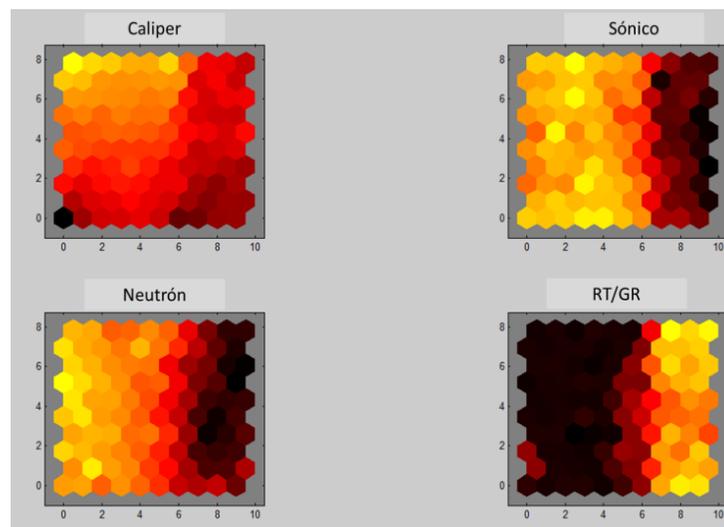


Fig. 4.38. Mapas de pesos para el modelo uno, por variable donde cada color representa el peso que le da una vecindad de neuronas a las entradas

La Fig. 4.38 muestra los mapas de peso por variable, en ella se observa que las variables del registro sónico y neutrón tienen formas similares, indicando alta correlación, lo cual se pudo comprobar por el cálculo de correlación mostrado en la matriz de dispersión de la Fig. A.19 del apéndice A.

La Fig. 4.39 muestra el mapa de éxito, donde se observan el número de datos que se encuentran en cada zona. En ella se pueden observar dos grupos principales y una serie de subdivisiones dentro de cada grupo (la línea punteada, muestra un par de ejemplos de subclases)

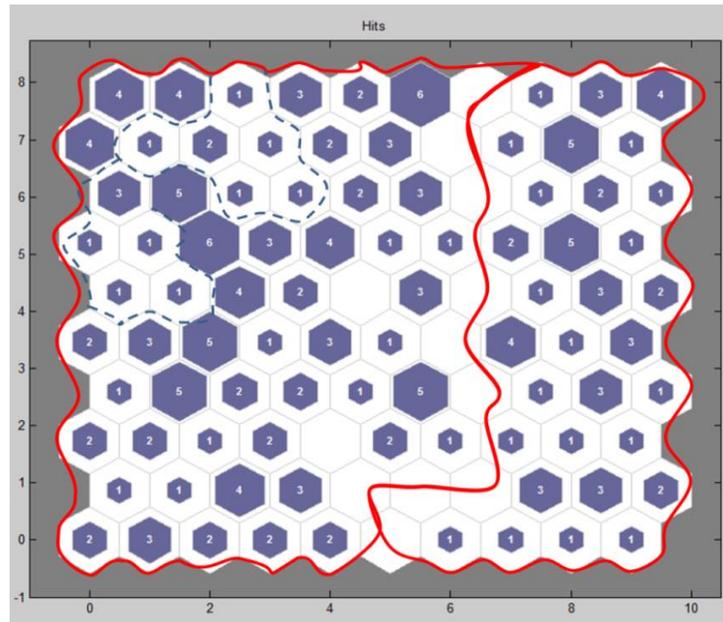


Fig. 4.39. Interpretación Mapa de éxitos del modelo uno, donde se observa dos zonas que pueden ser consideradas como grupos, muy diferente a los que pudiesen existir en otras zonas del mapa.

b) Entrenado con el modelo dos (medianamente clasificado)

Los parámetros utilizados fueron:

- Un mapa de 10 x 10 neuronas (100 neuronas), en una topología hexagonal.
- Se utilizaron únicamente los datos normalizados entre cero y uno.

La Fig. 4.40 muestra los pesos o distancia que tienen los datos para una determinada vecindad de neuronas. La Fig. 4.41 muestra los mapas de peso por variable, en ella se observa, que el grado de correlación que existe entre las variables, es bajo.

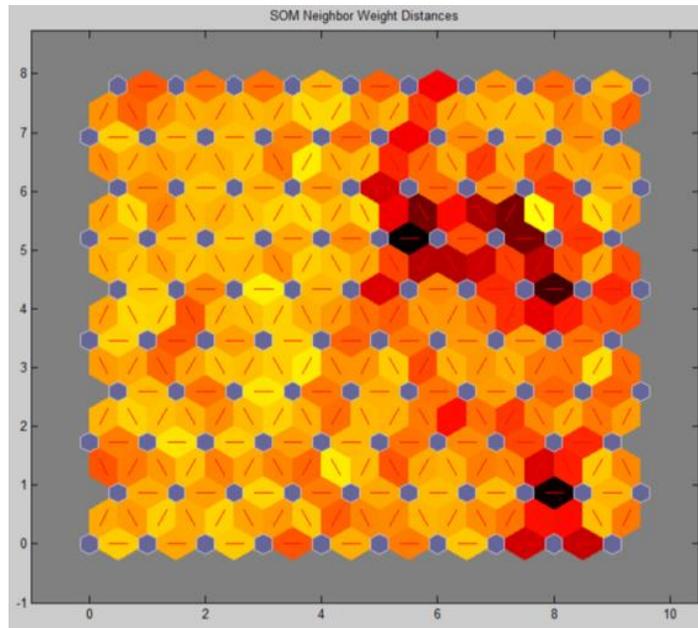


Fig. 4.40. Mapa de pesos para el modelo dos, donde cada color representa el peso que le da una vecindad de neuronas a las entradas

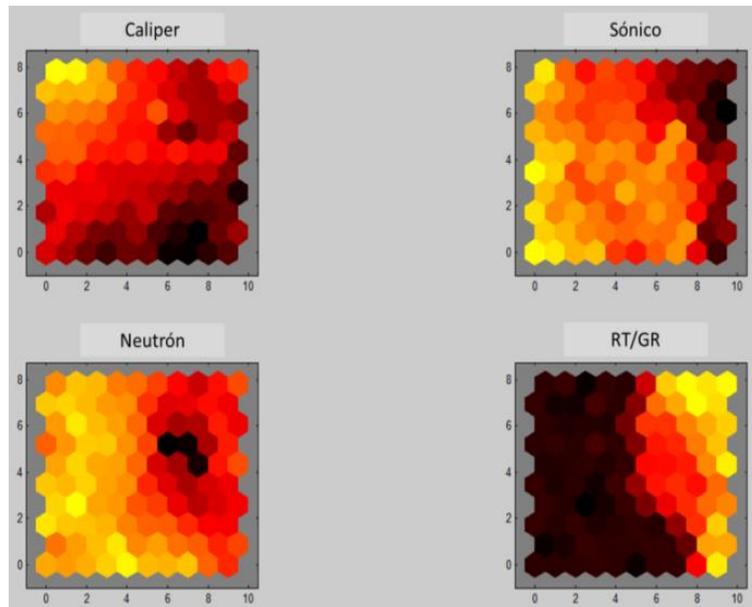


Fig. 4.41. Mapas de pesos para el modelo dos, por variable, donde cada color representa el peso que le da una vecindad de neuronas a las entradas

La Fig. 4.42 muestra una interpretación del mapa de éxito, donde se observan el número de datos que se encuentran en cada zona, identificando varias zonas bien definidas y algunas subclases (Por ejemplo las subclases marcadas con la línea punteada)

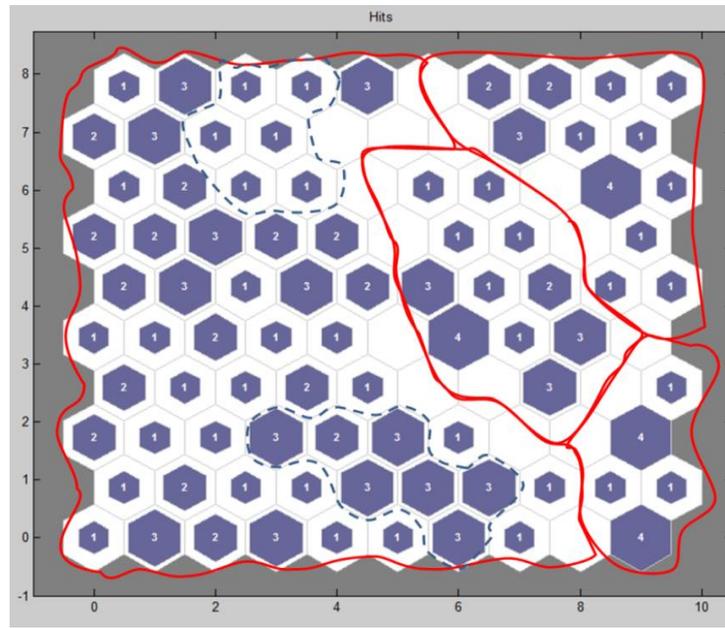


Fig. 4.42 Mapa de éxitos del modelo dos, donde se observa cuatro grupos y varias subclases en cada uno de ellos

c) Entrenado con el modelo tres (mal clasificado)

Los parámetros utilizados fueron:

- Un mapa de 10 x 10 neuronas (100 neuronas), en una topología hexagonal.
- Se utilizaron únicamente los datos normalizados entre cero y uno.

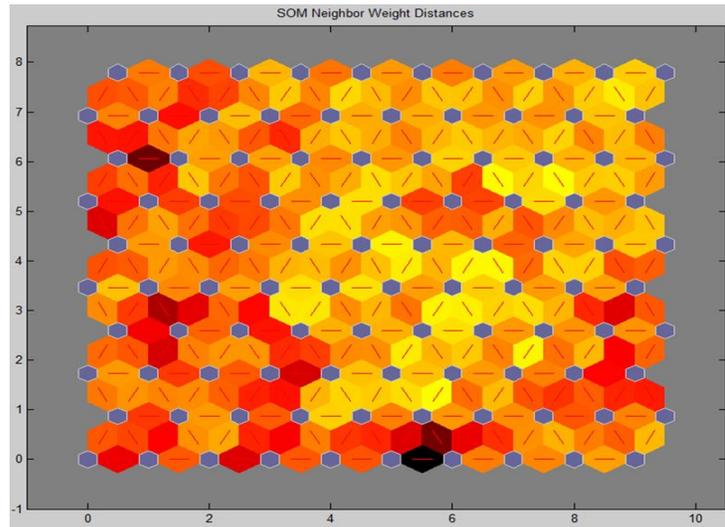


Fig. 4.43 Mapa de pesos para el modelo tres, donde cada color representa el peso que le da una vecindad de neuronas a las entradas

La Fig. 4.43 muestra los pesos o distancia que tienen los datos para una determinada vecindad de neuronas. La Fig. 4.44 muestra los mapas de peso por variable, en ella se observa nuevamente que las variables del registro sónico y neutrón son los que más grado de correlación tienen.

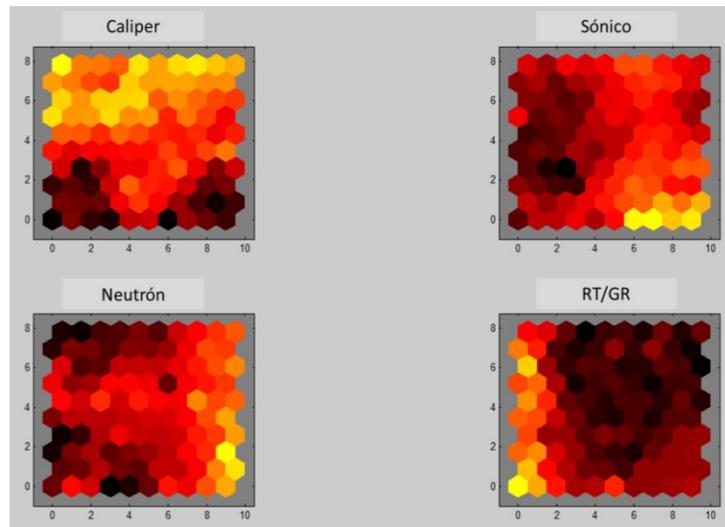


Fig. 4.44. Mapas de pesos para el modelo tres, por variable, donde cada color representa el peso que le da una vecindad de neuronas a las entradas

La Fig. 4.45 muestra el mapa de éxito, donde no se observa una separación muy clara de las clases, sin embargo, se pueden interpretar algunas subclases. Algunos ejemplos de estas subclases son definidas por la línea punteada de la Fig. 4.45

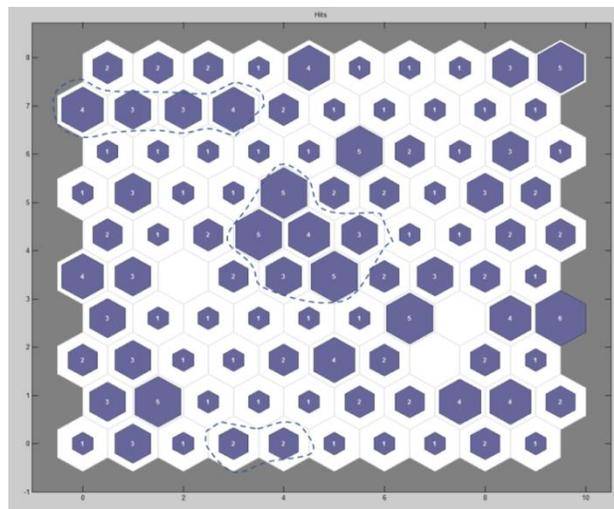


Fig. 4.45 Interpretación del mapa de éxitos del modelo tres, donde se observa dos zonas que pueden ser consideradas como grupos, muy diferente a los que pudiesen existir en otras zonas del mapa.

4.4.5.4 Resultados de la clasificación con Mapas Auto-organizados (SOM)

Los resultados de las estimaciones se muestran en la Fig. 4.46, donde se puede observar que los datos, por sí mismo sin un patrón o guía, se clasifican en más de tres clases.

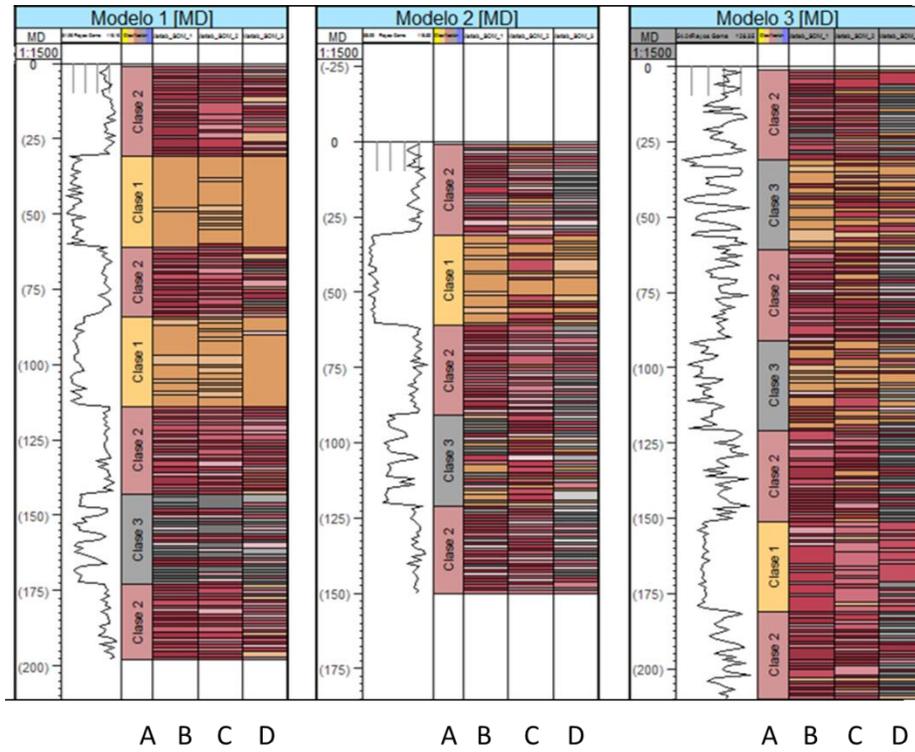


Fig. 4.46. Resultados de la clasificación usando mapas auto-organizados. A) Clasificación real, B) Entrenando con modelo uno, C) Entrenando con modelo dos, D) Entrenando con modelo tres

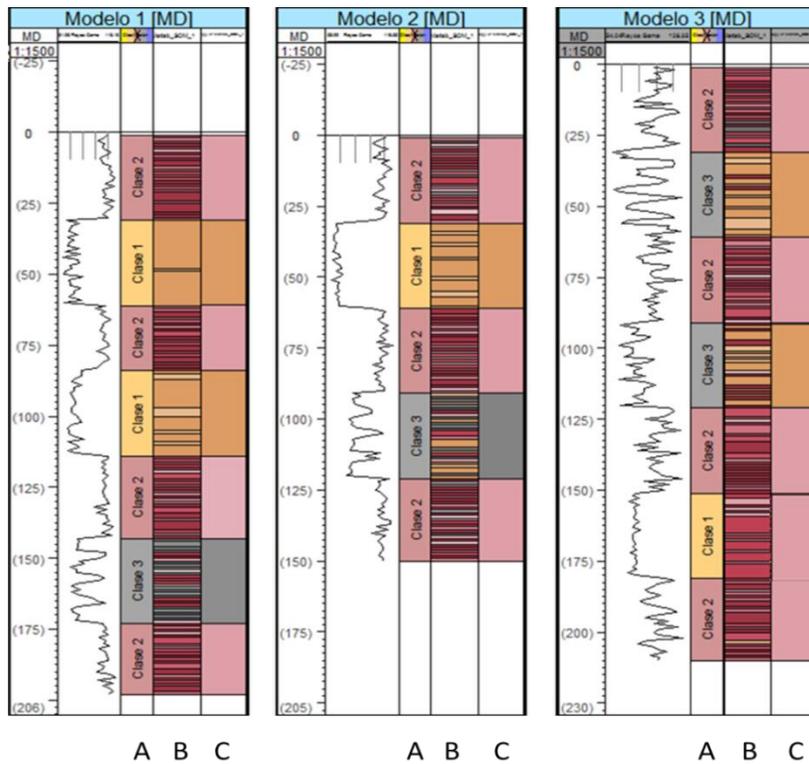


Fig. 4.47. Interpretación de la clasificación usando mapas auto-organizados. A) Clasificación real, B) Entrenando con modelo uno, C) Interpretación.

Para definir si cada subclase pertenece o no a una clase en específico, se utilizó como indicador el registro *RT/GR* (definido del análisis exploratorio de datos) y con ello se obtienen una serie de códigos, para cada subclase.

En general se puede observar que existe una zonificación, que coincide con la clasificación del modelo original; sin embargo, entrenando con el modelo tres y estimando los demás, se genera una clasificación de regular a mala. En la Fig. 4.47, se observa la interpretación de la clasificación obtenida de un mapa auto-organizado, entrenado con el modelo uno. Nuevamente se corroboró la dificultad de estimar el modelo tres, debido a la mala clasificación de los datos.

4.4.5.5 Máquinas Comité (CM)

Las máquinas comité se desarrollaron en Matlab, mediante el uso de códigos para redes de conjunto (ensemble neural network) y redes modulares (modular neural network) publicados por Kala R. (2013), editados, adaptados e integrados para los fines de este trabajo de tesis.

Se realizaron diversas pruebas, similares a las realizadas, con los demás modelos neuronales. Sin embargo, se utilizó el modelo uno, como modelo de entrenamiento final.

Se utilizaron tres módulos (tres litofacies), cada uno de ellos formado por tres redes neuronales perceptrón multicapa similares a la utilizada en la sección 4.4.5.1, entrenadas con los registros originales y las salidas de una red de conjunto (ensemble neural network). Cada red neuronal de conjunto está conformada por tres redes neuronales, similares a la utilizada en la sección 4.4.5.1 entrenadas con los datos originales. La diferencia entre las diferentes unidades individuales (MLP) que conforman la red modular o máquina comité (CM), fueron el número de neuronas para un mismo número de capas, los algoritmos de entrenamiento, la forma en que los datos son presentados a la red y los parámetros propios de los algoritmos de entrenamiento.

La Fig. 4.48, muestra los resultados finales de estas pruebas. Nuevamente se puede ver que el modelo tres, es el que tiene mayor error, ya que el conjunto de redes confunde la clase tres con la clase uno.

Se puede observar que el sistema modular de redes neuronales confunde la clase tres y la clase uno, en el modelo tres, ya que en el análisis exploratorio de datos, se puede observar que los datos son muy semejantes, a excepción del registro de rayos gama, sin embargo, se infiere que el modelo uno (con el que se entrena la red modular) no es el indicado para estimar el modelo tres, debido a la diferencia entre ellos (ver análisis exploratorio de datos del Apéndice dos), sin embargo se obtuvieron resultados muy similares a la estimación de este modelo con mapas auto-organizados.

4.4.5.6 Conclusiones del entrenamiento con redes neuronales.

El mayor problema es la similitud de los datos y la generalización de los mismos. Del análisis de exploratorio de datos, se pudo constatar la diferencia que existe entre los modelos uno y dos, con el modelo tres. Entrenando cualesquiera de los modelos neuronales ya sea con el modelo uno o el modelo dos, se puede estimar el otro con un grado de aproximación aceptable. El modelo tres es prácticamente diferente a los otros dos modelos, por lo que los métodos no supervisados, son los que mejor lo representan, sin embargo, es difícil para los diferentes algoritmos diferenciar la clase uno y tres, dada la naturaleza de los datos (la cual es un reflejo de sus características petrofísicas).

Para el caso de los mapas auto-organizados, los resultados pueden reflejar inconsistencia en el modelo tres, sin embargo, el análisis de *k* medias, permitió obtener una mejor clasificación que este modelo neuronal, por lo que se podría concluir que el modelo neuronal no es óptimo para clasificar el modelo tres.

En el caso de las redes perceptrón multicapa y máquinas comité, existe la necesidad de contar con una mayor cantidad (los datos son insuficientes), que sean representativos de la clase, además del hecho que

entrenando con el modelo uno ó dos. Para el modelo tres es necesario entrenar la red con un modelo que sea mucho más parecido al modelo tres.

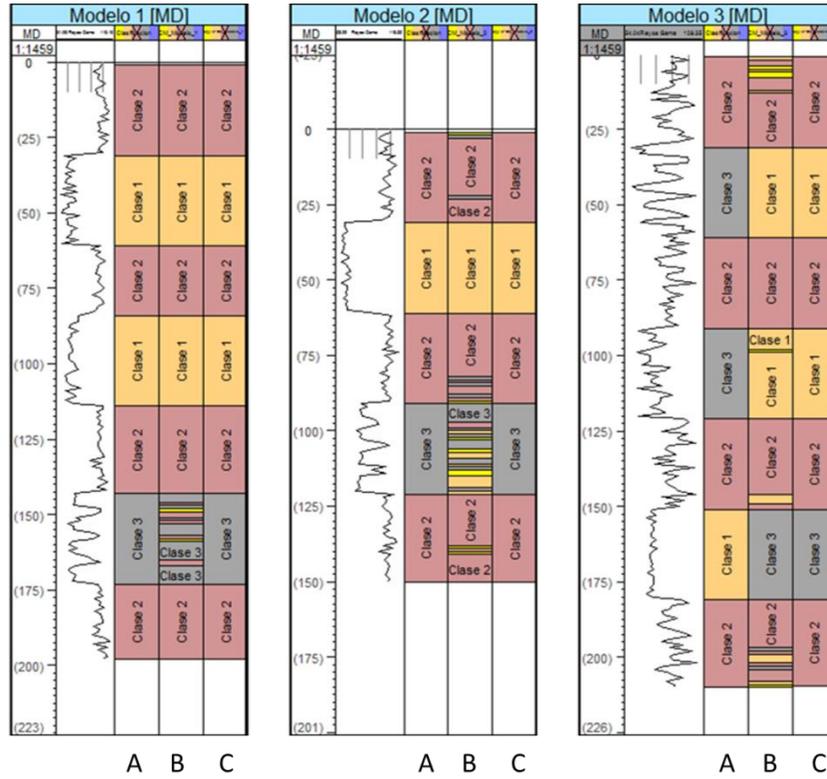


Fig. 4.48. Interpretación de la clasificación usando máquinas comité. A) Clasificación real, B) Entrenando con modelo uno, C) Interpretación

V. APLICACIÓN DE RED NEURONAL ARTIFICIAL A UN CASO DE ESTUDIO

En presente capítulo, se documenta la metodología propuesta en el capítulo anterior, aplicada a datos reales de registros geofísicos de pozo, en un arreglo de macropera de siete pozos y un pozo cercano a este arreglo. Los pozos se encuentran dentro de un campo perteneciente al proyecto Aceite Terciario del Golfo de PEMEX Exploración y Producción, en el paleocanal de Chicontepec. Los datos de pozo, son utilizados por medio del convenio que establece el Proyecto Fondos SENER-CONACYT¹² del Instituto Mexicano del Petróleo (No. 116606) y el oficio PEP-SGRT-GEP-480-2013, referente al permiso otorgado para la realización de este trabajo de tesis.

En este capítulo se muestran los resultados más relevantes de la metodología propuesta, con el objetivo de brindar una herramienta que apoye la caracterización de electrofacies, durante la etapa de conceptualización del modelo geológico y la etapa del poblado de propiedades petrofísicas, durante el proceso de caracterización estática de yacimientos.

Este trabajo de tesis se cuenta únicamente con registros geofísicos de pozos, para el entrenamiento de las redes neuronales. La interpretación manual de electrofacies en los registros de pozo, fueron interpretadas por personal experto de un agrupamiento multidisciplinario del Instituto Mexicano del Petróleo. Dichas interpretaciones serán el punto de partida para entrenar los sistemas supervisados y para analizar los resultados simulados mediante los diferentes modelos neuronales.

5.1 Antecedentes

El Paleocanal de Chicontepec se ha convertido en los últimos años en un área petrolera de vital importancia para México. Es el segundo campo petrolero más importante debido a sus importantes reservas estimadas. La compleja geología y la falta de tecnología complican la extracción de hidrocarburos generando una recuperación baja (Jiménez y Sandoval, 2009).

Las primeras manifestaciones de aceite dentro del paleocanal de Chicontepec, datan de 1926. Los yacimientos están constituidos por areniscas y lutitas de la edad del Eoceno Inferior- Paleoceno Superior que comprenden una columna aproximada de 1,000 m. de espesor bruto, extendida en una superficie aproximada de 625 Km² (Jiménez y Sandoval, 2009).

Geológicamente se localiza entre la porción sur de la cuenca de Tampico-Misantla y al occidente de la Plataforma de Tuxpan; esta cuenca se originó en el Paleoceno, al inicio de los levantamientos de la Sierra Madre Oriental (Orogenia Laramide) y fue afectada en su margen Oriental y Suroriental por una depresión de considerable magnitud denominado Paleocanal de Chicontepec. La sedimentación del Paleoceno se compone de turbiditas de ambiente nerítico externo a batial, formando complejos de abanicos y canales, constituida por arenas lenticulares con intercalaciones de lutitas. Durante este período sucedieron varios eventos de erosión y relleno de sedimentos. En el Eoceno Medio se observa una sedimentación arcillosa de ambiente batial (Jiménez y Sandoval, 2009).

La columna geológica investigada por los pozos comprende desde el Jurásico Inferior hasta el Oligoceno Superior. En el sureste del área, los yacimientos terciarios del Paleocanal de Chicontepec descansan discordantemente en el Jurásico Superior San Andrés; en la parte centro y Norte, se encuentran por encima del Cretácico Superior Méndez. El Paleoceno está representado por las formaciones Velasco, Chicontepec Inferior y Chicontepec Medio. El Eoceno Inferior, por las formaciones Aragón y Chicontepec Superior Canal. El Eoceno Medio por la Formación Guayabal y el Eoceno Superior por las formaciones Tantoyuca y Chapopote. En Chicontepec se tienen definidos tres plays productores: Play Chicontepec (Paleoceno-Eoceno Inferior), Play Chicontepec Canal (Eoceno Inferior Tardío) y Play Tantoyuca (Eoceno Superior) (Jiménez y Sandoval, 2009).

La Cuenca de Chicontepec se caracteriza por contener grandes espesores de unidades siliciclásticas de ambiente marino. El modelo geológico está definido como una serie de eventos múltiples, como depósitos de

¹² Proyecto Fondos CONACYT SENER "Inyección alternada de agua y gas (WAG) como sistema de recuperación mejorada"

abanicos submarinos, erosión y relleno. En la mayoría de las unidades genéticas existentes, se pueden destacar cuatro elementos del sistema de depósito del modelo de abanico: abanico proximal, medio y distal. Las facies del abanico medio son facies canalizadas o rellenos de canal (conglomerados y arenas, ver facies tipo B del modelo Galloway de la Fig. 5.1). Las facies del abanico proximal son canales y lóbulos (Flujos de escombros, arenas canalizadas, lóbulos de arena con intercalaciones de lutita y levas. Facies A, B, C, D y E de la Fig. 5.1). Las facies del abanico distal, se caracterizan por litofacies de arenas finas, arenas arcillosas, limolitas y lutitas.

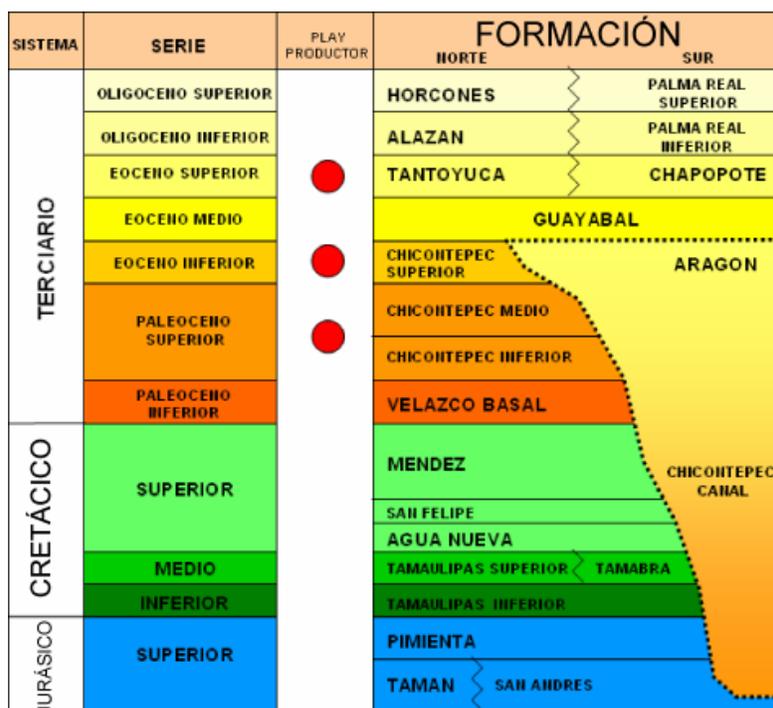


Fig. 5.1. Columna geológica del paleocanal de Chicontepec. Tomada de Jiménez y Sandoval (2009), modificada de Cuevas (2005)

5.2 Información disponible.

La información que se cuenta es de 8 pozos:

- Pozo 311
- Pozo 312
- Pozo 318
- Pozo 331
- Pozo 332
- Pozo 333
- Pozo 351
- Pozo 353

La distribución de los pozos se muestra en la Fig. 5.2, donde se observa el arreglo de los pozos y sus desviaciones. En un pozo desviado la interpretación de los registros de pozo tiene algunas complicaciones, por ejemplo los cálculos de espesor de los yacimientos, está en función de inclinación del pozo y el echado de la capas de roca. En ocasiones la desviación del pozo no permite que exista acoplamiento adecuado entre algunas herramientas de medición de los registros y la pared del pozo, generando errores de medición.

En este trabajo no se considera los efectos que pudiesen tener las desviaciones, ya que se considera que los registros han sido editados por un petrofísico experto.

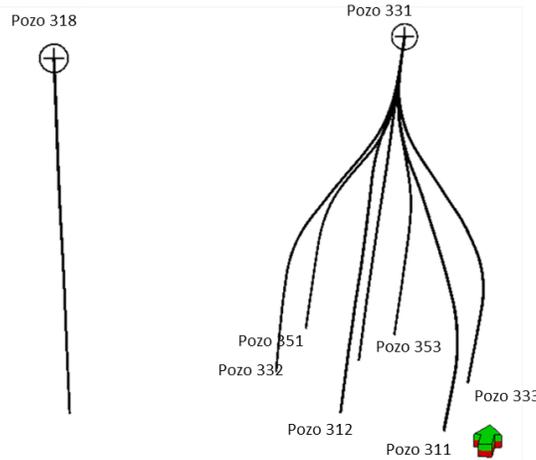


Fig. 5.2. Distribución de los pozos

Los datos de registros geofísicos presentan diferentes acrónimos, por lo que fue necesario buscar su equivalencia (Ya que para entrenar las redes debemos hacerlo con el mismo tipo de dato), la cual se resume en la Tabla 5.1

Tabla 5.1. Información disponible.

	Rayos Gama (GR)	Neutrón (NPHI)	Resistividad (RT)	Densidad (RHOB)	Caliper (CALI)	Sónico (DT)
Pozo 311	✓	✓	✓	✗	✓	✗
Pozo 312	✓	✓	✓	✓	✓	✓
Pozo 318	✓	✓	✓	✓	✓	✓
Pozo 331	✓	✓	✓	✓	✓	✓
Pozo 332	✓	✓	✓	✓	✓	✓
Pozo 333	✓	✓	✓	✓	✓	✓
Pozo 351	✓	✓	✓	✓	✓	✓
Pozo 353	✓	✓	✓	✓	✓	✓

Prácticamente los pozos poseen todos los a excepción del pozo 311 que no cuenta con el registro sónico ni de densidad. En este trabajo de tesis, se realizó el análisis y la clasificación automatizada de facies en el intervalo que comprendió la prueba de inyección de CO₂. En la Fig. 5.3 se muestra una sección de todos los pozos disponibles, en el intervalo de análisis, con los registros de rayos gama y resistividad, colgados desde la cima de la formación Chicontepec.

Estos registros fueron analizados por geólogos y petrofísicos expertos del Agrupamiento Multidisciplinario del Instituto Mexicano del Petróleo, para un proyecto de inyección de CO₂, donde interpretaron cinco facies de depósito (mediante asociación visual de los registros), cuatro diferentes litofacies y cuatro clases petrofísicas, las cuales están en función de las siguientes propiedades petrofísicas: volumen de arcilla, porosidad efectiva, permeabilidad y saturación de agua.

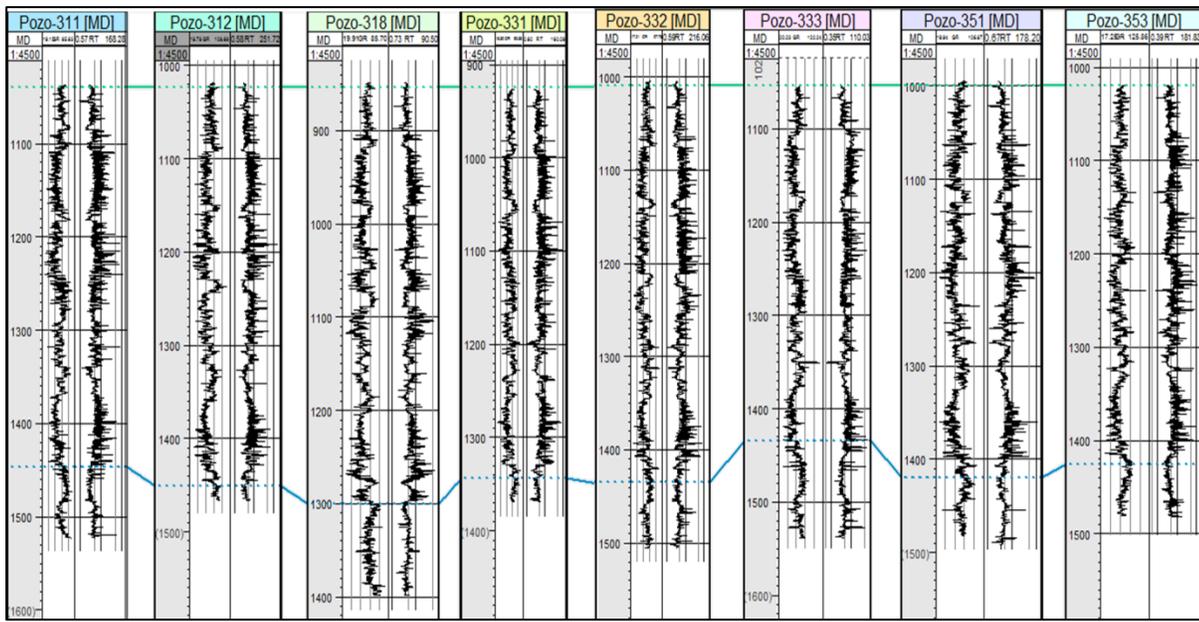


Fig. 5.3. Sección de los pozos utilizados

5.3 Clasificación Manual

La clasificación manual de facies realizadas por los expertos consto de los siguientes pasos (Casar et. Al., 2012):

- (1) Se parte de los pozos que cuentan con núcleos y los datos de registros de pozo más completos y confiables, que se encuentren en las áreas más representativas del yacimiento.
- (2) Las facies se definen e identifican en los registros de pozo y se ajustan con la información de núcleos, en función del patrón de apilamiento y la forma de la curva de los registros rayos gama y resistividad.
- (3) Se agrupan en un número reducido de litotipos o litofacies, en función de los valores de corte principalmente del registro rayos gama y resistividad.
- (4) La información es validada y analizada en función de los gráficos neutrón-densidad-rayos gama, neutrón-sónico-rayos gama, sónico-densidad-neutrón y gráficos de picket (resistividad-porosidad)
- (5) Cuando las propiedades petrofísicas del yacimiento no están directamente controladas por las facies o litotipos, se puede usar un enfoque alternativo en términos de clases petrofísicas (agrupamiento por propiedades petrofísicas).
- (6) El esquema final de clasificación es extendido al resto de los pozos.

Para fines prácticos, dentro del flujo de trabajo del modelado geológico y petrofísico de un yacimiento, se decidió utilizar en este trabajo, una clasificación de electrofacies, basada en la litología (litofacies), en lugar de clasificaciones basadas en patrones de apilamiento o propiedades petrofísicas.

Las electrofacies estimadas son:

- | | | | |
|----|--------------------|---|---|
| a. | Arenisca |  | Rango de valores del registro rayos gama: 0-50 |
| b. | Arenisca arcillosa |  | Rango de valores del registro rayos gama: 50-58 |
| c. | Lutita Arenosa |  | Rango de valores del registro rayos gama: 58-70 |
| d. | Lutita |  | Rango de valores del registro rayos gama: > 70 |

Esta clasificación se muestra en la sección de pozos de la Fig. 5.4, donde se observa que el pozo 318, es el único pozo que no cuenta con esta clasificación.

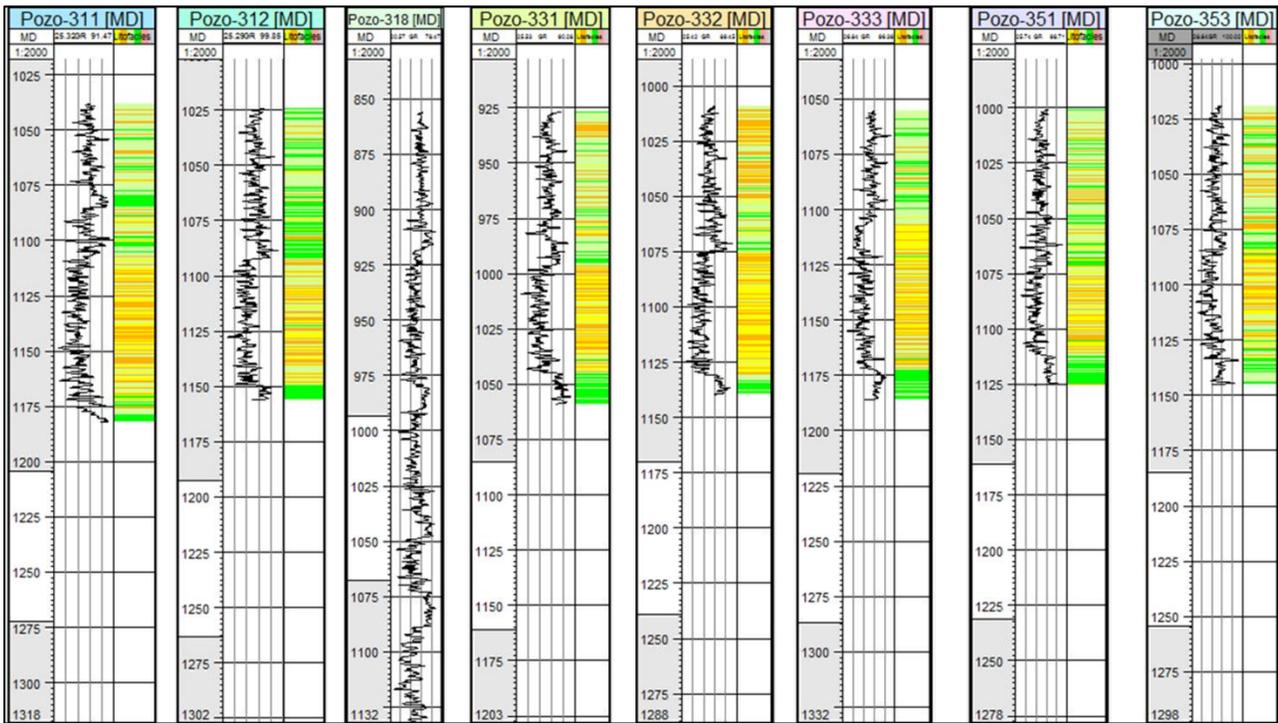


Fig. 5.4. Facies determinadas por especialistas, en el intervalo de inyección de CO₂

5.4 Aplicación de la metodología

Se aplicó el flujo de trabajo propuesto a los datos de registros de los pozos del campo en estudio (sección 5.3), seleccionando los registros de pozo que tengan mejor consistencia estadística y representen mejor la clasificación.

En este trabajo se propone nuevamente el uso de la relación RT/GR , debido a su utilidad durante el entrenamiento de los datos sintéticos del capítulo cuatro. La derivada de los rayos gama con respecto a la profundidad, no fue propuesta en este caso debido a que no aportó información valiosa durante la clasificación de registros sintéticos.

5.4.1 Censo y revisión de la información (control de calidad).

Para esta etapa, es importante recordar que los datos utilizados fueron editados, corregidos (efectos del pozo) y analizados por un petrofísico experto, por lo que se considera que los datos utilizados están en óptimas condiciones. A pesar de lo anterior, es importante realizar un análisis visual rápido de las curvas, tratar de revisar si existen problemas con ellas y generar interpretaciones rápidas de las mismas, con el fin de validar el dato correctamente. En este caso se revisaron los datos de las curvas, con rápidas interpretaciones visuales,

observando que los datos tienen comportamiento ruidoso (ruido de alta frecuencia). Una opción viable es que los registros pueden ser suavizados, siempre y cuando no se afecte la naturaleza del dato. En el caso de los pozos 311 y 318 se observa derrumbes, donde el registro caliper, difiere de los derrumbes del pozo 333 y 351, para el mismo nivel estratigráfico. Se observa además que el pozo 351 tiene ausencia de datos en los registros de rayos gama, caliper, densidad, sónico y neutrón (A partir de 1490 m aprox.) y como consecuencia la relación RT/GR también. Con respecto a la clasificación existen zonas donde hay variabilidad de un punto a otro, lo cual generalmente se debe a errores puntuales

5.4.2 Análisis exploratorio de datos

El análisis exploratorio de datos se realizó en su totalidad con el lenguaje de programación R. El objetivo es analizar el comportamiento estadístico de los datos de los pozos en estudio. El análisis exploratorio de datos (*AED*), fue generado siguiendo los pasos de la metodología del capítulo tres, sin embargo, de la experiencia adquirida al trabajar con datos sintéticos, el análisis exploratorio de datos mostrado en el Apéndice B, fue enfocado para caracterizar el comportamiento estadístico de las cuatro electrofacies, bajo el siguiente esquema:

- a) Con todos los datos de todos los pozos, se calcularon los estadísticos básicos y los gráficos estadísticos, por electrofacies.
- b) El análisis de correlaciones por pozo, mediante gráficos de dispersión

Los análisis generales por pozo y por pozo-electrofacie, se realizaron para entender el comportamiento de los datos, sin embargo, no se muestran en este trabajo.

5.4.2.1 Análisis por electrofacie

Los resultados del análisis exploratorio de datos por electrofacie mostrados en el Apéndice B, permiten analizar los valores de los registros que caracterizan a cada electrofacie. El análisis por electrofacie permite ver las diferencias o similitudes entre electrofacies, para un registro en específico.

Estos análisis se realizaron considerando los datos de todos los pozos, a excepción del pozo 318, el cual solo cuenta con una clasificación de 3 clases petrofísicas (mostradas en el gráfico de dispersión de la Fig. B.22), la cual no es considerada para este análisis.

a) Rayos Gama (*GR*).

Las Tablas B.1, B.2, B.3 y B.4, muestran los resúmenes de los estadísticos analizados por electrofacie, donde las medias y medianas del registro de rayos gama (indicador del contenido de arcilla) se encuentran en un rango de valores muy amplio (aprox. 45 a 75 API's), con desviaciones estándar similares y con valores de límites de electrofacie muy cercanos entre sí (valores mínimos y máximos de cada electrofacie), lo cual permite inferir que para este registro las electrofacies tenderán a parecerse. De los análisis solo por pozo (gráficos de caja de la Fig. B.19) puede verse que las distribuciones de los datos en cada pozo son relativamente homogéneas y simétricas y que sus valores medios no son tan parecidos pero se encuentran cercanos. Se observa además, que el pozo 318 difiere de los demás pozos. Para el análisis solo por pozo no se muestran los histogramas; sin embargo estos no presentaban multimodalidad.

Las Figs. B.1, B.2, B.3 y B.4, muestran los histogramas del registro de rayos gama por pozo, donde se observa el amplio rango de valores que se puede encontrar para cada electrofacie, principalmente las electrofacies híbridas (Arena arcillosa y lutita arenosa), lo cual puede significar que existen diferencias entre los valores de una electrofacie en los diferentes pozos (Esto se observa en el análisis exploratorio de datos por pozo y por pozo-electrofacie, aunque no es muy evidente). Las electrofacies más puras (arena limpia y lutita franca), tienen distribuciones asimétricas, por lo que sus valores tienden a parecerse a las electrofacies híbridas, por lo que se espera, que un algoritmo clasificador pueda confundir los valores la electrofacie uno con la electrofacie dos y la

electrofacie cuatro con la electrofacie tres. De los gráficos de caja de la Fig. B.5, se observa la diferencia del rango de valores medios y los rangos intercuartílicos, para las cuatro electrofacies.

En los gráficos cuantil-cuantil de las Figs. B.6, B.7, B.8 y B.9, se observa que existe un conjunto de valores que se aproximan a un comportamiento normal (en las cuatro electrofacies), lo cual refleja que un número de datos tiene un comportamiento estable y aleatorio alrededor de la media, por lo que los puntos que no se comportan de manera normal, podrían ser valores atípicos, por lo que es necesario analizarlos y entender si corresponden a un comportamiento diferente que no ha sido muestreado o son errores asociados a las mediciones. Los gráficos cuantil-cuantil del registro de rayos gama, permiten identificar las posibles modas (la modas se identifican cada vez que la curva corta la línea que representa el comportamiento normal).

Dentro de esta metodología es necesario analizar, mediante el análisis exploratorio de datos, el efecto de los valores atípicos y cualquier transformación realizada a los datos. Los resultados de estos análisis, no son presentados en este trabajo, sin embargo a partir de ellos se tomó la decisión de trabajar directamente con los datos originales.

b) Resistividad (*RT*).

En las Tablas B.1, B.2, B.3 y B.4, se observa, que no existe un contraste claro entre las facies, a excepción de la electrofacie uno, en la cual su valor promedio de resistividad es ligeramente mayor que las otras tres electrofacies, como se muestra en los gráficos de caja de la Fig. B.5, en donde también se muestra que los rangos intercuartílicos están muy cercanos entre sí, lo cual permite suponer que el registro de resistividad es un discriminante de la electrofacie uno y que el contenido de arcilla confundirá al clasificador usando este registro.

Los histogramas de las Figs. B.1, B.2, B.3 y B.4, permiten observar la superposición entre las electrofacies y el efecto que tienen los valores atípicos en ellas (muy probablemente, serán mal clasificados). Esta variable es de las más complejas para utilizar como parámetro de clasificación, debido a que la resistividad no es solo función de la litología, sino del contenido de fluidos que contenga la roca, la porosidad y el grado de compactación.

En los gráficos cuantil-cuantil de las Figs. B.6, B.7, B.8 y B.9, se observa que el comportamiento difiere del normal (en las cuatro electrofacies) y permite identificar la presencia de valores atípicos (observar los gráficos de caja de los histogramas correspondientes a cada electrofacie).

Para tratar de entender el comportamiento de los datos se analizaron los datos sin valores atípicos y con transformación logarítmica (no se muestra este análisis). La distribución mejora debido a que la naturaleza del registro de resistividad es log normal. No se considera alterar el dato, debido a que dentro del flujo existen otras transformaciones (normalizaciones, componentes principales, etc.).

c) Caliper (*CAL*).

De los resultados mostrados en las Tablas B.1, B.2, B.3 y B.4, y los gráficos de caja de las Figs. B.5, se puede identificar un contraste debido a su contenido arcilloso (el registro caliper determina el tamaño del agujero, que por sus características de permeabilidad genera enjarre de pozo y en formaciones arcillosas, que generalmente son más deleznable e inestables, produciendo derrumbes), por lo que existen mayores diferencias usando este registros para discriminar las electrofacies tres y cuatro.

Los histogramas de las Figs. B.1, B.2, B.3 y B.4, permiten observar comportamientos más normales, con valores más próximos a la media para las electrofacies uno y dos, a diferencia de las electrofacies tres y cuatro, que tienen mayor desviación estándar y por lo tanto un rango mayor de valores para cada una de estas electrofacies. Sin embargo, por lo parecidos de sus medias (Fig. B.5), es posible que el clasificador confunda las electrofacies más arenosas.

En los gráficos cuantil-cuantil de las Figs. B.6, B.7, B.8 y B.9, se observa que el comportamiento difiere del normal (en las cuatro electrofacies) y permiten identificar la presencia de valores atípicos. A este registro también le fue realizado el análisis sin datos atípicos.

d) Densidad (DEN).

Los resúmenes de los estadísticos para este registro, se muestran en las Tablas B.1, B.2, B.3 y B.4, se puede mencionar que los valores promedios para las cuatro electrofacies son prácticamente iguales, por lo que no será discriminador desde el punto de vista numérico. Sin embargo este es uno de los registros principales para realizar inferencias litológicas y mineralógicas, de manera manual durante una evaluación petrofísica.

Los histogramas de las Figs. B.1, B.2, B.3 y B.4, muestran que la distribución de los datos de densidad se encuentra dentro de rangos de valores muy similares entre sí, tendiendo un comportamiento más asimétrico conforme el contenido de arcillosidad aumenta. Se sabe que el registro de densidad responde a la matriz y al fluido. Las electrofacies uno y cuatro son las que mayor dispersión muestran, con un rango amplio de valores para definir esta dos electrofacies, lo cual permite concluir (en caso de que la clasificación de electrofacies sea consistente) que se tendrán arenas y lutitas, con comportamientos de compactación y contenido de fluidos diferentes. De los gráficos de caja de la Fig. B.5, se observan la diferencia del rango de valores medios y los rangos intercuartílicos para las cuatro electrofacies, donde la media de la electrofacie uno, es la que más se separa de las demás medias, sin embargo, esa diferencia no es significativa.

En los gráficos cuantil-cuantil de las Figs. B.6, B.7, B.8 y B.9, se observa que al igual que los registros de rayos gama, una parte de los datos tiene un comportamiento normal.

Para tratar de entender el comportamiento de los datos se graficaron los histogramas suprimiendo los valores atípicos y realizando una transformación logarítmica. Los resultados no se muestran en este trabajo y al igual que en las variables anteriores, se decidió no modificar el dato original.

e) Sónico (DT).

Sus estadísticos se muestran en las Tablas B.1, B.2, B.3 y B.4, donde se observa un comportamiento similar al registro de rayos gama, con la diferencia que los límites de las electrofacies son ligeramente más claros (pero existe superposición).

Los histogramas de las Figs. B.1, B.2, B.3 y B.4, son más simétricos en las electrofacies uno y dos, que en las electrofacies tres y cuatro. Las desviaciones estándares son similares. Los rangos de valores son amplios en cada electrofacie y sus límites se superponen. El registro sónico responde principalmente a la matriz, pero al ser de ondas compresionales, también responde al fluido. De los gráficos de caja de la Fig. B.5, se observa la diferencia del rango de valores medios y los rangos intercuartílicos, para las cuatro electrofacies. Este registro es consistente con la clasificación propuesta (al igual que el registro de rayos gama) y es un buen discriminador de las electrofacies.

En los gráficos cuantil-cuantil de las Figs. B.6, B.7, B.8 y B.9, se observa que el registro sónico es el que más se aproxima a una distribución normal, debido a que este registro está afectado por ruido de alta frecuencia típico del proceso de adquisición, por lo que los valores atípicos se deben principalmente a este fenómeno.

Se analizaron los datos eliminando los valores atípicos y realizando transformaciones logarítmicas, los resultados no se muestran en este trabajo, sin embargo, se puede concluir que eliminar los valores atípicos no mejora mucho la distribución de los datos en los histogramas, debido a que ya presentaban una buena distribución (Figs. B.1, B.2, B.3 y B.4). Se decidió por tanto utilizar los datos sin modificación (considerando que se deben generar más transformaciones como normalización de los datos o en componentes principales), ya que del análisis de sintéticos del capítulo cuatro, se observa que las redes neuronales tienen un grado de tolerancia al ruido aleatorio.

e) Neutrón (NPH).

Las Tablas B.1, B.2, B.3 y B.4, muestran los estadísticos para este registro, donde se observa que existen diferencias ligeramente claras entre los valores medios de cada una de las electrofacies.

Los histogramas de las Figs. B.1, B.2, B.3 y B.4, tienen un comportamiento similar al registro sísmico, siendo más simétricos en las electrofacies uno y dos, que en las electrofacies tres y cuatro. Las desviaciones estándares son similares. Los rangos de valores son amplios en cada electrofacie y se superponen entre las diferentes electrofacies. El registro de neutrones responde principalmente al contenido de fluidos, por lo que si se supone que el fluido es diferente en las arenas que en las lutitas (considerando hidrocarburos solamente), se tiene un parámetro discriminante de clasificación con este registro. De los gráficos de caja de la Fig. B.5, se observa la diferencia del rango de valores medios y los rangos intercuartílicos, para las cuatro electrofacies, con un comportamiento muy similar al sísmico.

En los gráficos cuantil-cuantil de las Figs. B.6, B.7, B.8 y B.9, se observa que el registro neutrón al igual que el registro sísmico se aproxima a una distribución normal, por lo que los problemas de adquisición y/o correcciones ambientales están presentes en el dato.

Nuevamente se analizaron los datos sin la presencia de valores atípicos y usando transformaciones de tipo logarítmico (no mostradas en este trabajo), nuevamente se decidió no modificar los datos.

5.4.2.2 Análisis de gráficos de caja (boxplots) por pozo

De la Fig. B.11, se puede observar de manera general que los pozos presentan el mismo comportamiento en los registros con respecto a sus valores medios y rangos intercuartílicos, a excepción del pozo 3.18, principalmente en los registros caliper, rayos gama y sísmico, que tienen comportamientos que se alejan de la media global, ya sea porque presenta condiciones geológicas, litológicas o físicas diferentes, o porque no fue normalizado en conjunto con los demás pozos durante el proceso de edición de datos realizados por el petrofísico.

5.4.2.3 Análisis de gráficos de dispersión (gráficos de correlación o scatter plots)

La Fig. B.10 muestra los gráficos de dispersión de todos los pozos y todas sus variables, poniendo en colores el tipo de electrofacie con la que un determinado punto ha sido clasificado de forma manual. Se puede mencionar que las electrofacies se encuentran regularmente dispersas en grupos (clústeres) bivariados, no muy bien definidos, ya que se superponen unas con otras. Se observa también que una determinada electrofacie, presenta subgrupos en algunos casos, lo cual puede estar asociado a la inconsistencia de la clasificación en los diferentes pozos (es decir que la electrofacie uno en un pozo no es igual a la electrofacie uno en otro) o que algunas subelectrofacies en realidad son electrofacies que no están lo suficientemente bien representadas. No existe un par de registros, que permita visualizar una separación de los agrupamientos clara, lo cual intuye que no hay un registro clave que permita discriminar entre una litología y otra.

Los gráficos B.12, B.13, B.15, B.16, B.17, B.18, y B.19, muestran los gráficos de dispersión de los pozos 311, 312, 331, 332, 333, 351 y 353, respectivamente, con el fin de mostrar el comportamiento de cada electrofacie en cada pozo y la consistencia de la clasificación. El comportamiento es muy similar al de la Fig. B.10, sin embargo, los agrupamientos se definen mejor (evitando el efecto de inconsistencia que existe en la clasificación de pozos), pero no obstante el efecto de superposición de electrofacies está latente en esos gráficos. El gráfico de dispersión del pozo 318 de la Fig. B.14, es diferente y utiliza un criterio de clasificación basado en clases petrofísicas (ya que no cuenta con una clasificación manual de electrofacies), observando un comportamiento similar a la clasificación de electrofacies, donde prácticamente la clase petrofísica 3 pasa desapercibida. Este ejercicio solo demuestra el grado de correlación de las variables.

5.4.2.4 Conclusiones del análisis exploratorio de datos.

Del análisis exploratorio, se puede concluir, que la clasificación de electrofacies propuesta presenta un grado de consistencia aceptable, sin embargo, los amplios rangos de valores que caracterizan a cada electrofacie, hace que exista superposición entre electrofacies. Del análisis de todos los pozos por electrofacies, junto con el análisis por pozo-electrofacie, permite observar que una misma electrofacie tendrá comportamientos muy parecidos, pero no iguales de un pozo a otro, lo cual deja prever que las redes neuronales de tipo supervisado

(MLP y CM), podrán tener un cierto grado de asertividad al momento de clasificar un pozo entrenando la red neuronal con otro. Para los modelos no supervisados (SOM), es mucho más recomendable entrenar con cada uno de los pozos a clasificar, que elegir un pozo, entrenar la red con él y clasificar los demás (esta conclusión fue corroborada con el estudio de registros sintéticos). Del análisis de los gráficos de correlación por pozo, se puede ver que los registros rayos gama, sónico y neutrón, son los que mejor discriminan la clasificación propuesta, los cuales pueden ser apoyados por los registros de resistividad y caliper. En este caso la relación RT/GR , tiene un comportamiento similar al registro de resistividad.

El pozo 311, no tiene todo el conjunto de registros, por lo que su entrenamiento será un caso particular. El pozo 318 no tiene una clasificación de electrofacies, por lo que será estimado usando los demás pozos. Del análisis por electrofacie (junto con los análisis por pozo y por pozo-electrofacie), se puede elegir entre los pozos 312, 331, 332, 333, 351 y 353, para el análisis, lo recomendable es realizar pruebas durante el entrenamiento de las redes neuronales.

5.4.3 Análisis de agrupamiento (clústeres)

El análisis de agrupamientos se realizó en su totalidad por medio del lenguaje R.

b) Análisis de Dendrogramas

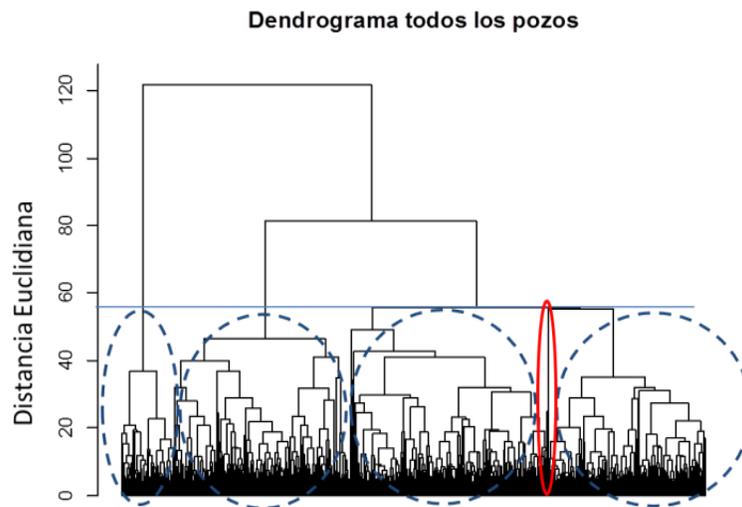


Fig. 5.5 Dendrograma para todos los pozos

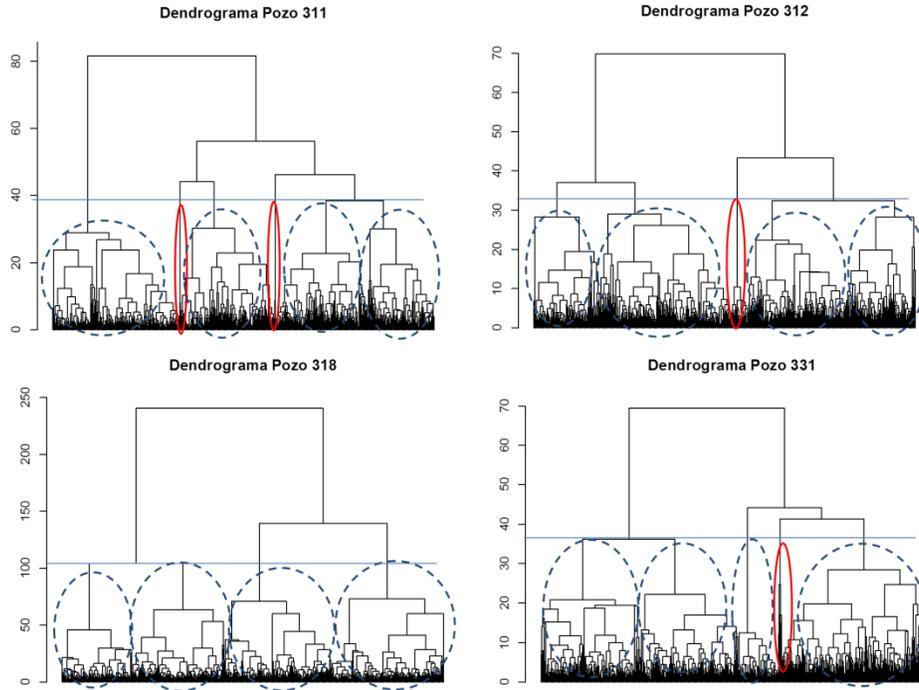


Fig. 5.6. Dendrograma de pozos 311, 312, 318 y 331.

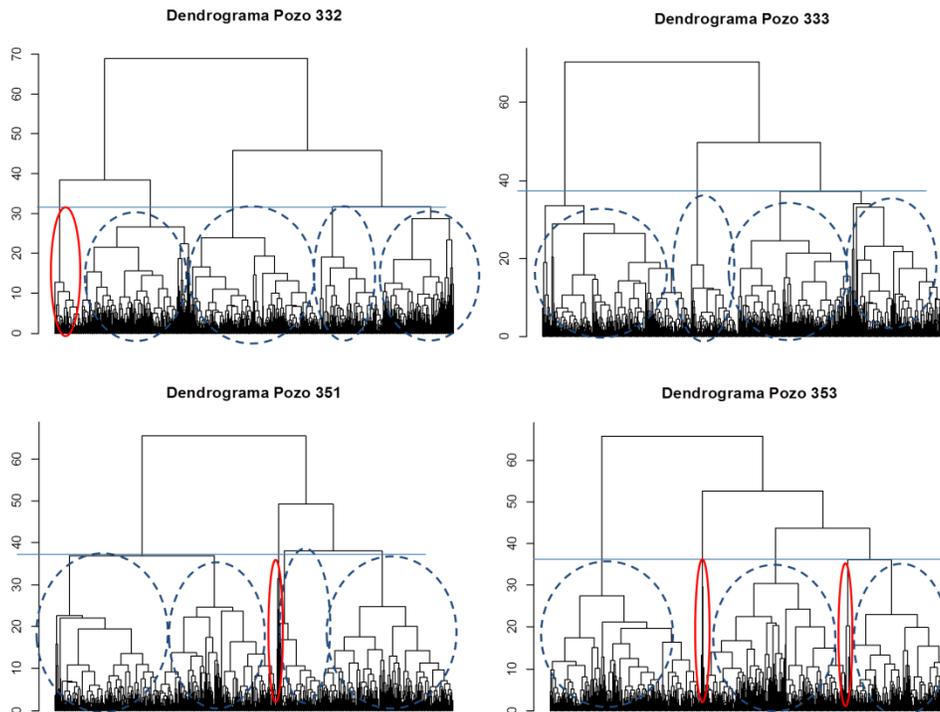


Fig. 5.7. Dendrograma de pozos 332, 333, 351 y 353

Para el análisis de dendrogramas se utilizó como criterio de similitud la distancia euclidiana, generando los dendrogramas para todo el conjunto de pozos y por pozo. La Fig. 5.5, muestra el dendrograma para todo el conjunto de pozos, utilizando los registros de rayos gama, resistividad, caliper, sónico y neutrón; donde se puede observar que si trazamos una línea perpendicular al eje de la distancia euclidiana, alrededor de un valor de 57, se observa que se forman cuatro grupos claros (óvalos de líneas punteadas) y un cuarto, pobremente representado (óvalo de línea continua). El mismo análisis se realizó para los ocho pozos de forma individual. Se puede apreciar el mismo comportamiento en las Figs. 5.6 y 5.7, en donde al intentar definir cuatro litofacies (como en los pozos 318 y 333), se forman un mayor número de electrofacie, como por ejemplo en los pozos 312,331, 332, 351 y 353, en donde se definen cinco litofacies o el pozo 311 en donde se definen seis litofacies.

c) K-medias

Apoyados de los dendrogramas antes analizados, se procede a utilizar el algoritmo de las k medias para la clasificación, suponiendo 4 electrofacies y utilizando los registros rayos gama, resistividad, caliper, neutrón y sónico. Los resultados de dicha clasificación de cada pozo se muestran en las Figs. 5.8 y 5.9, en donde, se muestran los registros de rayos gama en el carril uno, las litofacies estimadas por los expertos del IMP en el carril dos y las electrofacies estimadas mediante el algoritmo de las k-medias, en el carril tres (a excepción del pozo 318, que se muestran en el carril dos), bajo la siguiente relación de colores:

Electrofacie - 1 Arenisca		Grupo - 1	
Electrofacie - 2 Arenisca arcillosa		Grupo - 2	
Electrofacie - 3 Lutita arenosa		Grupo - 3	
Electrofacie - 4 Lutita		Grupo - 4	

De estos resultados puede observarse que la clasificación es regular, con respecto a la clasificación original. La electrofacie tres es asociada al grupo cuatro, la electrofacie cuatro al grupo 2, la electrofacie dos al grupo 3 y la electrofacie uno con el grupo 1.

La Fig. 5.10, muestra gráficos de puntos, para todos los pozos, donde se observan el número de observaciones contra el valor de la variables elegidas (relación *GR*, *RT*, *CALI*, *DT* y *NPHI*), con el objeto de analizar los rangos de valores que comprende cada grupo, para cada una de las variables. Desde el punto de vista univariado, se puede observar que el método de las k medias esta siendo consistente para clasificar los datos, por rango de valores.

Las Figs. 5.11 y 5.12 muestran los gráficos de dispersión de los registros *GR*, *RT*, *CALI*, *DT*, *NPHI* (los mismos de la Fig. B.10), con los centroides de cada grupo y el centroide global. Se puede observar que de manera bivariada existe una superposición de las electrofacies y no es factible identificar los diferentes grupos (clústeres). Se observa además que los centroides de los grupos (clústeres), están muy juntos uno del otro y muy proximos al centroide global, lo que implica mala clasificación de los datos, ya que estos, en terminos de la distancia, son muy parecidos.

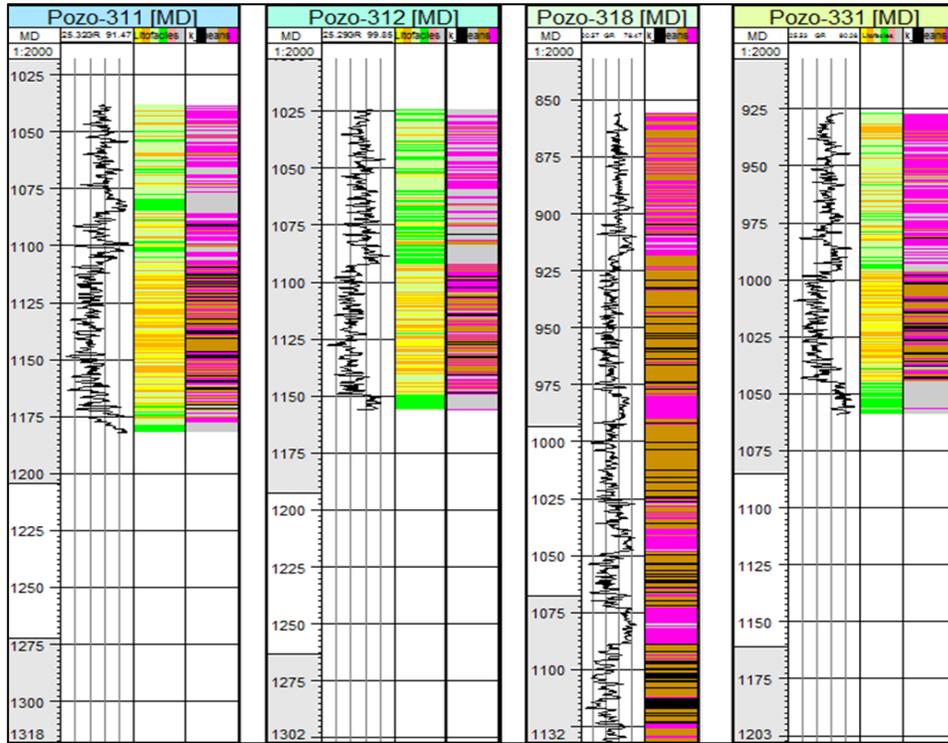


Fig. 5.8. Resultado de la clasificación por el método de k medias, de los pozos 311, 312, 318 y 331. (Carril 1, registro de rayos gama, carril dos la clasificación propuesta por los expertos, carril tres, clasificación mediante el algoritmo de k medias)

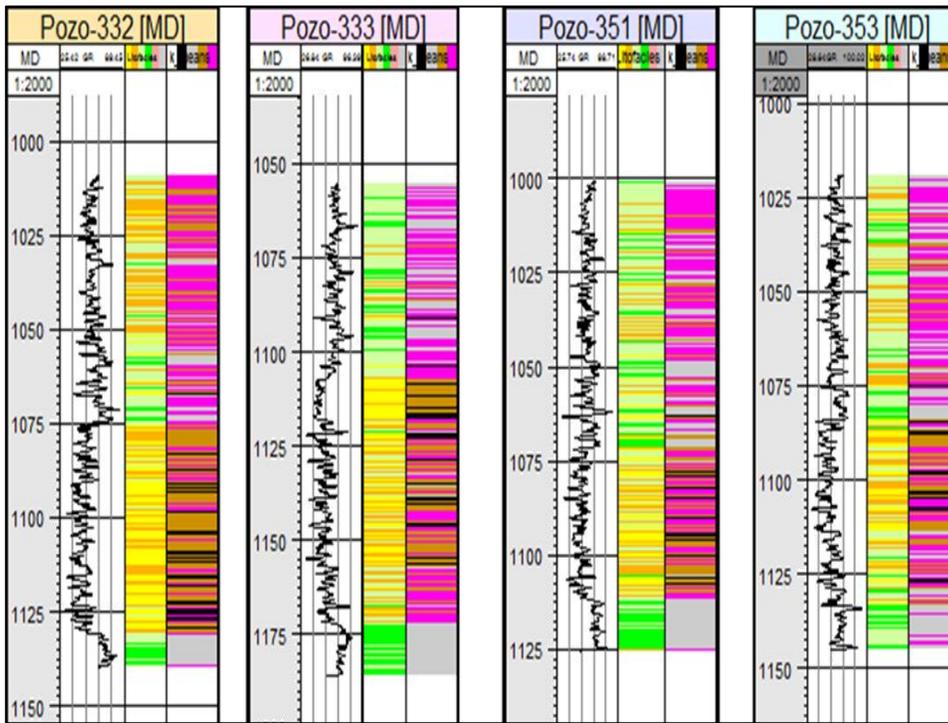


Fig. 5.9. Resultado de la clasificación por el método de k medias, de los pozos 332, 333, 351 y 353. (Carril 1, registro de rayos gama, carril dos la clasificación propuesta por los expertos, carril tres, clasificación mediante el algoritmo de k medias)

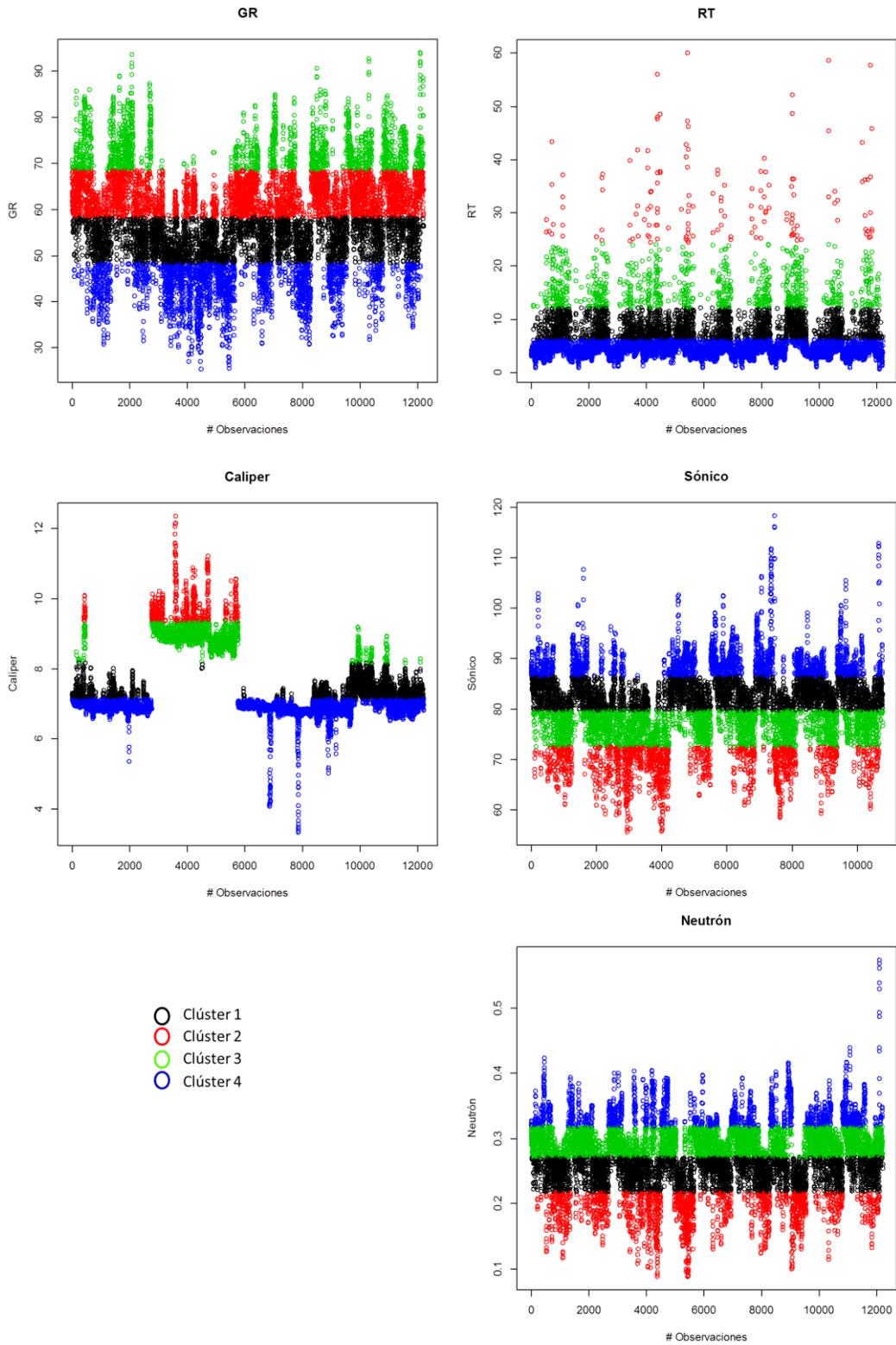


Fig. 5.10. Gráfico de puntos modelo uno, resultado de la clasificación por el método de k medias.

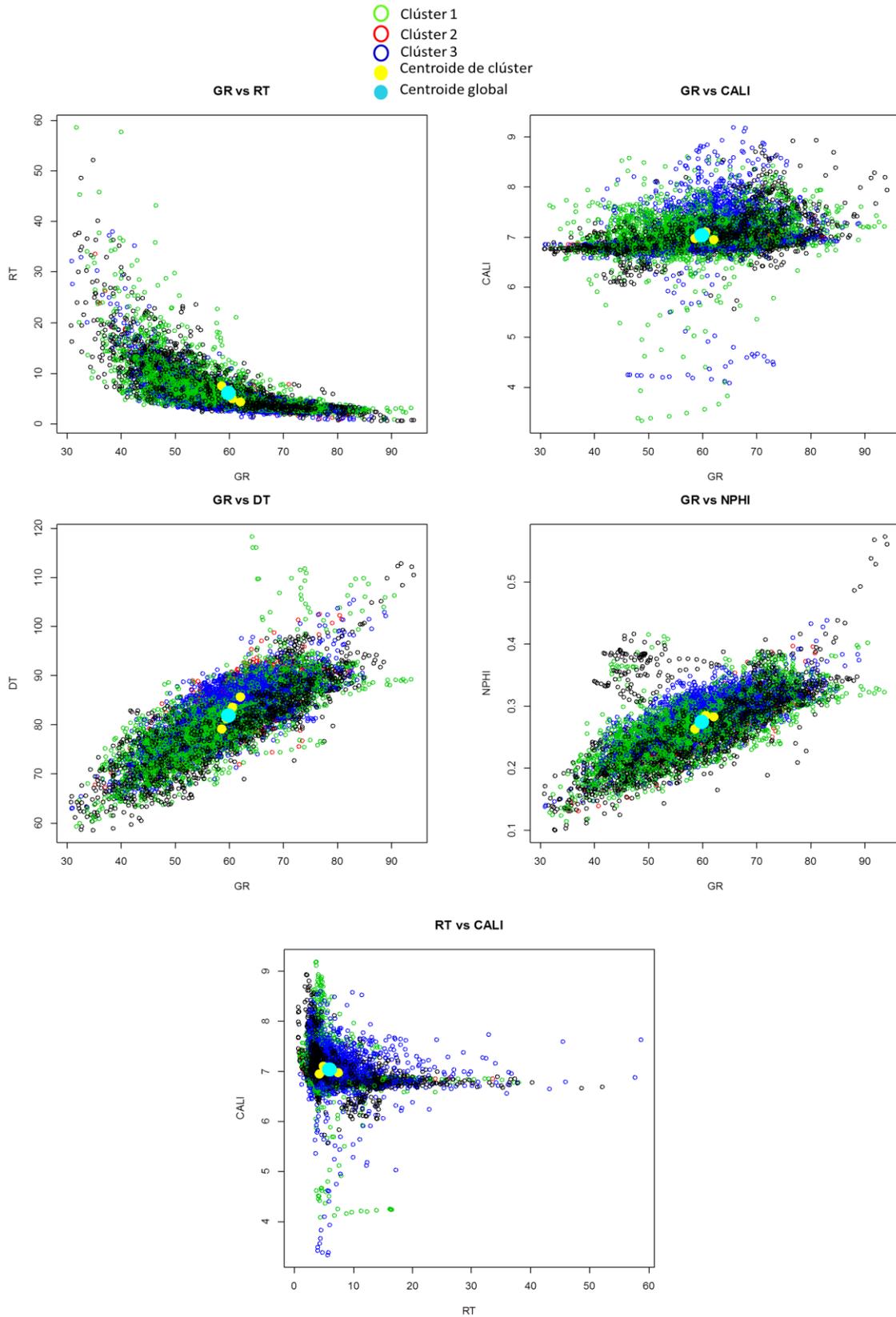


Fig. 5.11. Gráficos de dispersión, para todos los pozos de los registros *GR*, *RT*, *CALI*, *DT* Y *NPHI*

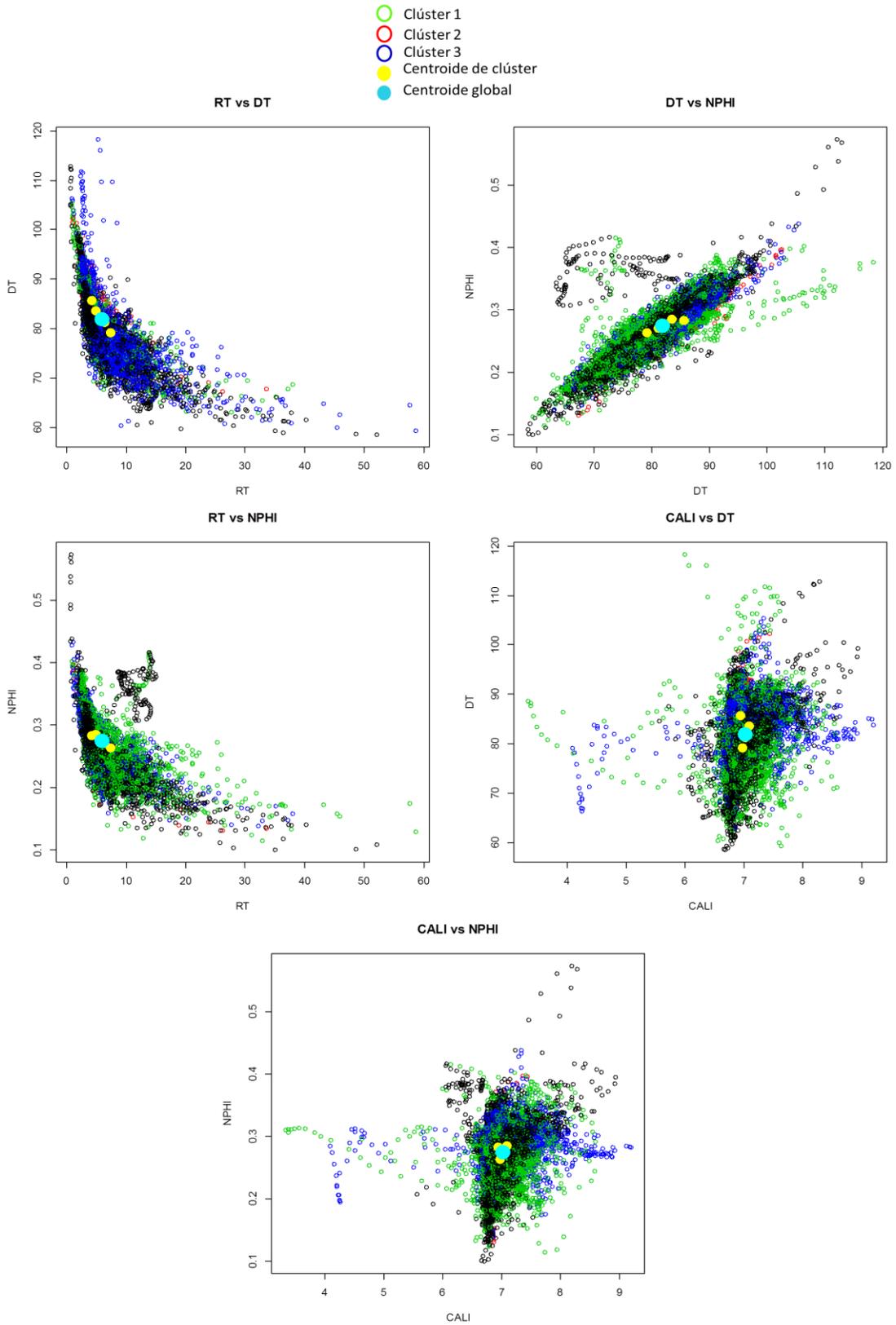


Fig. 5.12. Gráficos de dispersión, para todos los pozos de los registros *GR*, *RT*, *CALI*, *DT* y *NPHI*

5.4.3.1 Conclusión del análisis de agrupamiento

Del análisis de agrupamiento (clústeres), se puede mencionar que la cantidad de variables, utilizadas en la clasificación es un factor fundamental, durante los procesos de clasificación. De este análisis junto con el análisis exploratorio de datos se puede determinar que las variables que mejor clasifican los datos, son los registros rayos gama, neutrón y sónico, apoyados del registro de resistividad y caliper. En el siguiente apartado de este capítulo, se realizará el análisis de componentes principales, con estas variables, para determinar las variables que son más óptimas para entrenar las redes neuronales.

La clasificación por medio del algoritmo de las k-medias, tiene problemas al clasificar las litofacies más arenosas (1 y 2) y menor problema con las arcillosas (3 y 4).

5.4.4 Análisis de componentes principales y selección de variables.

El análisis de componentes principales, es aplicable cuando al analizar los coeficientes de correlación de Pearson, tienden a aproximarse a 1 o -1, lo que significa mayor grado de correlación lineal. El objetivo como se mencionó en los capítulos 3 y 4, es determinar las variables que proporcionen mayor información al clasificador (en este caso una red neuronal) de manera óptima. Del análisis exploratorio de datos y el análisis de agrupamiento (clústeres) de determino que los registros rayos gama, sónico y neutrón, serán la base para el proceso de clasificación.

El análisis se realizó para todos los pozos en conjunto y para cada pozo individual, sin embargo debido a la similitud de los resultados del análisis exploratorio y el análisis de componentes principales por pozo, en este capítulo se muestra únicamente el análisis global de todos los pozos en conjunto y el análisis de biplots, tanto para todo el conjunto de pozos, así como el análisis individual para cada pozo.

d) Todos los pozos en conjunto

La Tabla 5.2, muestra las variables que están más relacionadas con cada una de las componentes, siendo el registro sónico, la que mayor relación tiene con la componente 1 y el caliper con la componente 2. Los registros *GR*, *NPHI* y *RT*, tienen prácticamente los mismos pesos para la componente 1 y componente 2, principalmente aportando la mayor contribución a la componente 1. La Tabla 5.3 muestra los valores propios para cada componente, siendo hasta la componente 2 los valores que llegan a 1, y según la regla de Kaiser, serían suficientes para representar todas las variables. Lo anterior se muestra también en los gráficos de sedimentación de la Fig. 5.13

Tabla 5.2. Relación de las variables con cada una de las componentes, para los modelos 1,2 y 3

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
<i>CALI</i>	-0.05751072	0.95246401	-0.08806697	-0.28120451	0.05170159
<i>DT</i>	0.52887802	-0.02492302	-0.17893278	0.01589279	0.8290943
<i>GR</i>	0.49939521	-0.18902527	-0.10890571	-0.76949664	-0.33299901
<i>NPHI</i>	0.49541418	0.16560352	-0.50284882	0.53803728	-0.4298828
<i>RT</i>	-0.47132767	-0.17039974	-0.83397235	-0.19764173	0.11933993

Tabla 5.3. Valores propios para cada componente

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
3.185784	1.0654886	0.3588789	0.2491974	0.140651

La Tabla 5.4, muestra la proporción de varianza que explica cada componente, por ejemplo la componente 1 representa el 63.7 % de la varianza total, la componente 2 el 21.3 % y así sucesivamente.

Tabla 5.4. Relación de las variables con cada una de las componentes.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Desviación estándar	1.7848765	1.0322251	0.59906504	0.49919675	0.3750347
Proporción de varianza	0.6371568	0.2130977	0.07177578	0.04983948	0.0281302
Proporción acumulada	0.6371568	0.8502545	0.92203032	0.9718698	1

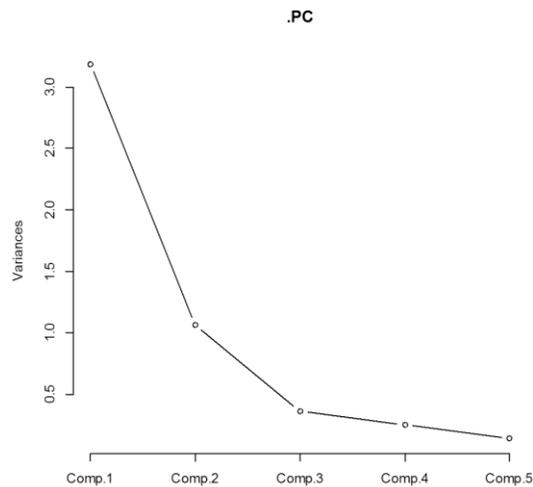


Fig. 5.13. Gráficos de sedimentación, para los análisis de modelos en conjunto y para los modelos 1, 2 y 3 de manera individual.

e) Análisis de componentes principales y análisis de agrupamiento (clústeres).

Mediante el algoritmo de componentes principales y el algoritmo de k-medias (definiendo cuatro grupos o clústeres), se generaron biplots, para el análisis utilizando todos los pozos (Fig. 5.14) y cada uno de los pozos de manera individual (Fig. 5.15). Los biplots muestran la relación que guardan las dos primeras componentes principales y las variables *GR*, *RT*, *CALI*, *DT* y *NPHI* variables relacionadas con ellas. Para este análisis, se realizó el análisis de agrupamiento (clústeres) por k-medias, definiendo cuatro grupos (clústeres) y graficando los biplots.

Los puntos más cercanos representan puntos con correlaciones similares para las componentes uno y dos, así como valores similares de las variables utilizadas (representadas como vectores), para cada análisis. Los puntos que caen cerca de los origen de los vectores, se aproxima a la media de cada variable. Los puntos más alejados de los vectores indican valores bajos, los puntos cercanos a los vectores representan valores más altos, por lo que se pueden detectar valores atípicos (outliers). A partir de la interpretación de los puntos se puede analizar diferentes grupos (clústeres).

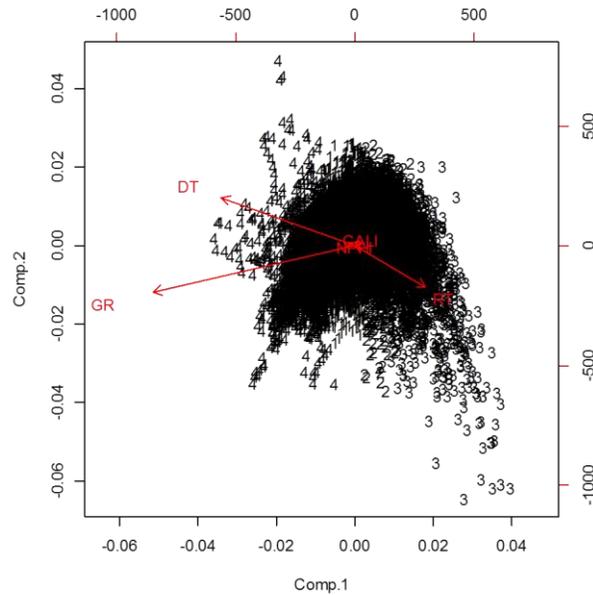


Fig. 5.14. Biplot para el análisis de componentes principales y de agrupamiento (clústeres), usando todo los pozos.

La magnitud y dirección de los vectores, están relacionados con el grado de correlación multivariado entre las variables, propiamente dicho la longitud está relacionada con la variabilidad y el ángulos que forman dos vectores, con la covariabilidad (correlación). La dirección está relacionada con el grado de correlación de los componentes principales y principalmente marca tendencias.

En la Fig. 5.14, se puede observar el grado de correlación entre las variables, siendo *DT*, *RT* y *GR*, las que mayor varianza (longitudes mayores) y covarianza (ángulos mayores), lo que supondría que son las variables que mejor responden a los cambios asociados a las facies, dentro del conjunto de datos de todos los pozos de manera global.

En la Figs. 5.15, se realiza el mismo análisis para cada uno de los pozos de manera individual, observando que para cada uno de ellos las variables *DT*, *GR* y *RT*, presentan el mismo comportamiento que en la Fig. 5.11, con diferentes valores de varianza para las variables en los pozos.

5.4.4.1 Conclusiones del análisis de componentes principales.

Como conclusiones de este análisis, se tiene que los registros que mejor definieron los grupos al realizar la clasificación por medio el algoritmo de k-medias (*GR*, *DT* y *RT*), son los que muestran mayor grado de varianza, por lo que influyen más en las componentes principales. En teoría, las componentes que deberían clasificar mejor los datos son las que contienen la mayor cantidad de información y representan la mayor variabilidad de los datos; sin embargo, los grupos (clústeres) se definen más utilizando el análisis de componentes principales con los registros *GR*, *RT*, *CALI*, *DT* y *NPHI*; por lo que se proponen este conjunto de registros para el entrenamiento con redes neuronales.

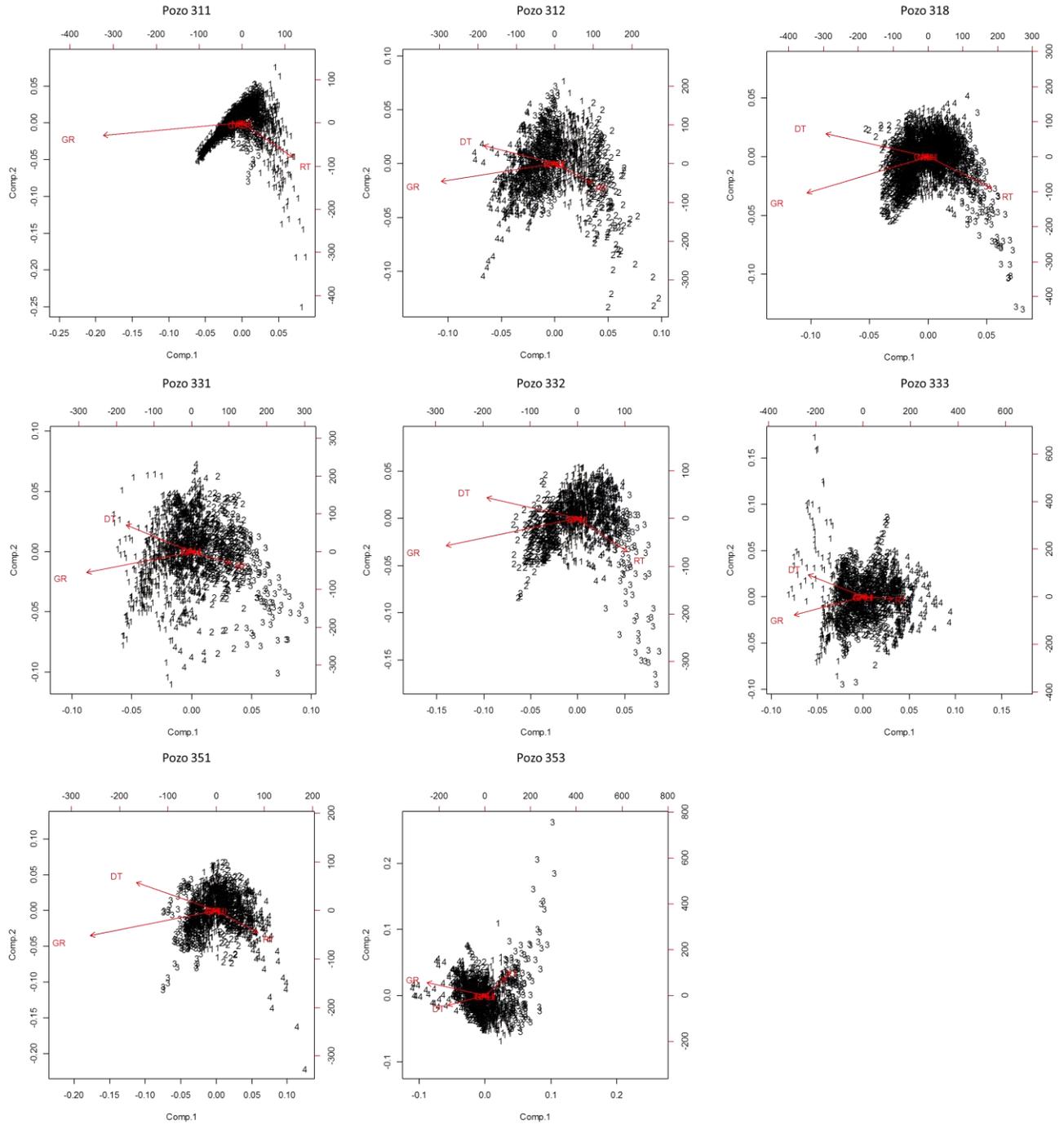


Fig. 5.15. Biplots para el análisis de componentes principales y de agrupamiento (clústeres), para cada uno de los pozos

5.4.5 Redes neuronales

La clasificación de los pozos propuestos se realizó mediante programación de los diferentes modelos neuronales: perceptrón multicapa (*MPL*), mapas auto-organizados (*SOM*) y las máquinas comitá (*CM*), en lenguaje M de MATLAB y su Neural Toolbox.

5.4.5.1 Redes perceptrón multicapa (MLP)

De los análisis mostrados en las secciones anteriores de este capítulo, se decidió utilizar los registros *GR*, *CALI*, *DT*, *RT* y *NPHI* del pozo 333, ya que este pozo es el que presenta los valores más estables (en términos de la media, la varianza y el rango intercuartílico).

Se generaron diversas pruebas, modificando los parámetros de entrenamiento principalmente el número de capas y neuronas, así como pruebas cambiando las variables (registros) del conjunto de datos de entrada, considerando el análisis exploratorio de datos.

Los parámetros utilizados fueron:

- Tres capas de ocultas con 10, 5 y 2 neuronas para la primera, segunda y tercer capa oculta respectivamente.
- Las funciones de transferencia fueron “tansig” para la primera capa, “logsig” para la segunda y tercer capa oculta, así como para la capa de salida.
- Los datos de entrada fueron los registros de rayos gama, resistividad, neutrón, sónico y caliper, elegidos del análisis exploratorio de datos y del análisis de componentes principales.
- Se normalizaron los datos, en función de la media y la desviación estándar, utilizando el comando de Matlab “*mapstd*” (Ec. 3.18)
- Se realizó una segunda prueba entrenando la red utilizando las dos primeras componentes principales.
- Se dividieron los datos de entrenamiento, validación y prueba, en 60%, 20% y 20%, respectivamente.

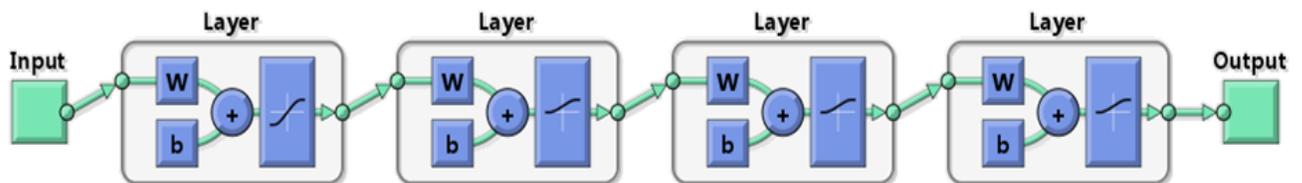


Fig. 5.16. Modelo de red neuronal

Las Figs. 5.16, 5.17 y 5.18, resumen el proceso de entrenamiento y estimación. En las Figs. 5.16 y 5.17, se observa el grado de rendimiento de la red neuronal para los diferentes conjuntos (entrenamiento, validación y prueba), obteniendo coeficientes de correlación de Pearson alrededor de 0.8.

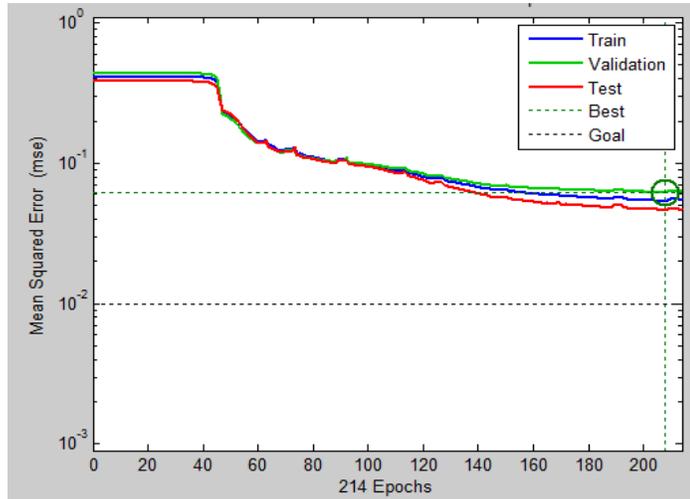


Fig. 5.17. Gráfico del monitoreo del error cuadrático medio, durante cada época (epoch)

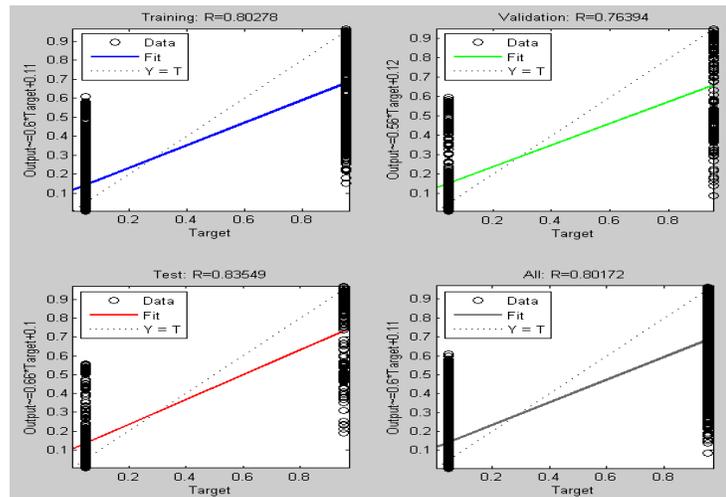
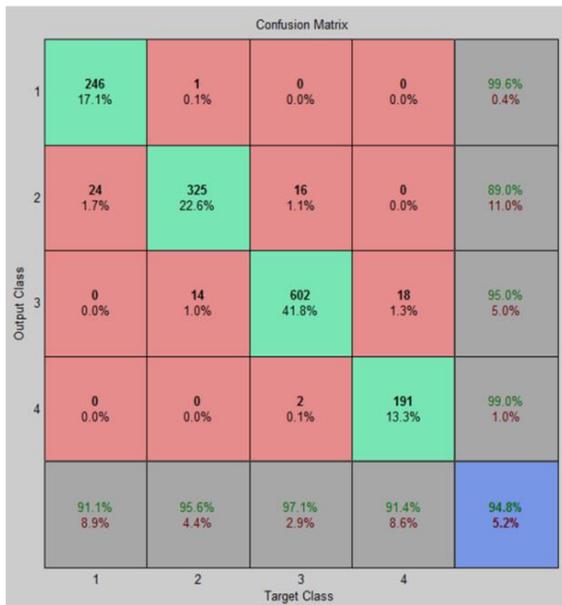


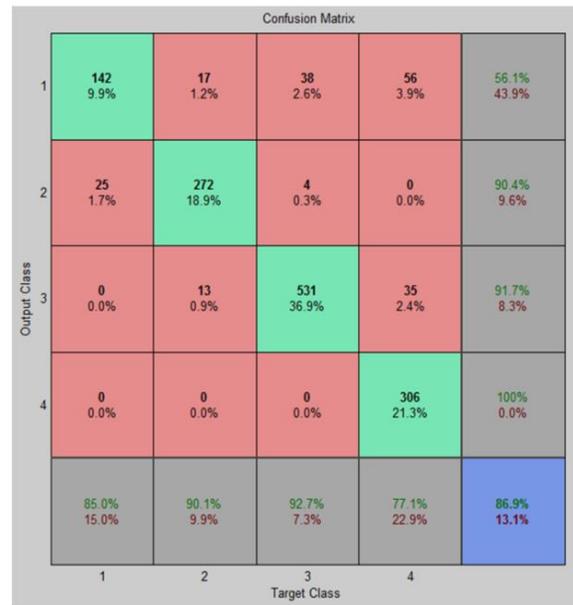
Fig. 5.18. Cuantificación del coeficiente de Pearson y gráfico de regresión lineal de los objetivos y las salidas para los diferentes conjuntos de datos (de izquierda a derecha y de arriba abajo, conjunto de entrenamiento, conjunto de validación, conjunto de prueba y una salida global)

De las matrices de confusión de las Figs. 5.19 y 5.20, se observa que los porcentajes de los datos bien clasificados con respecto a las facies propuestas por los expertos, es de alrededor del 85 al 93%, siendo el pozo 312, el que menor grado de reconocimiento tuvo.

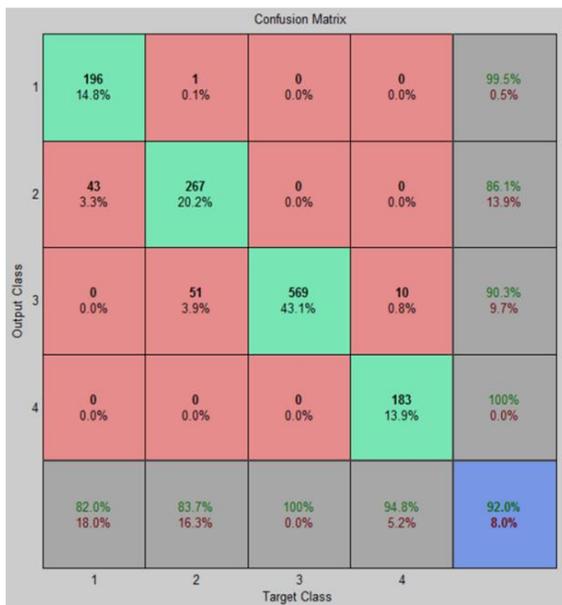
Pozo 311



Pozo 312



Pozo 331



Pozo 332

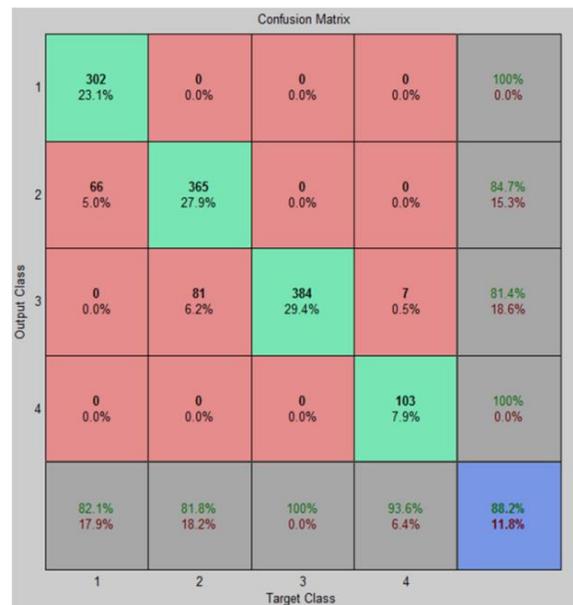


Fig. 5.19. Matrices de confusión, de la clasificación de los pozos 311, 312, 331 y 332.

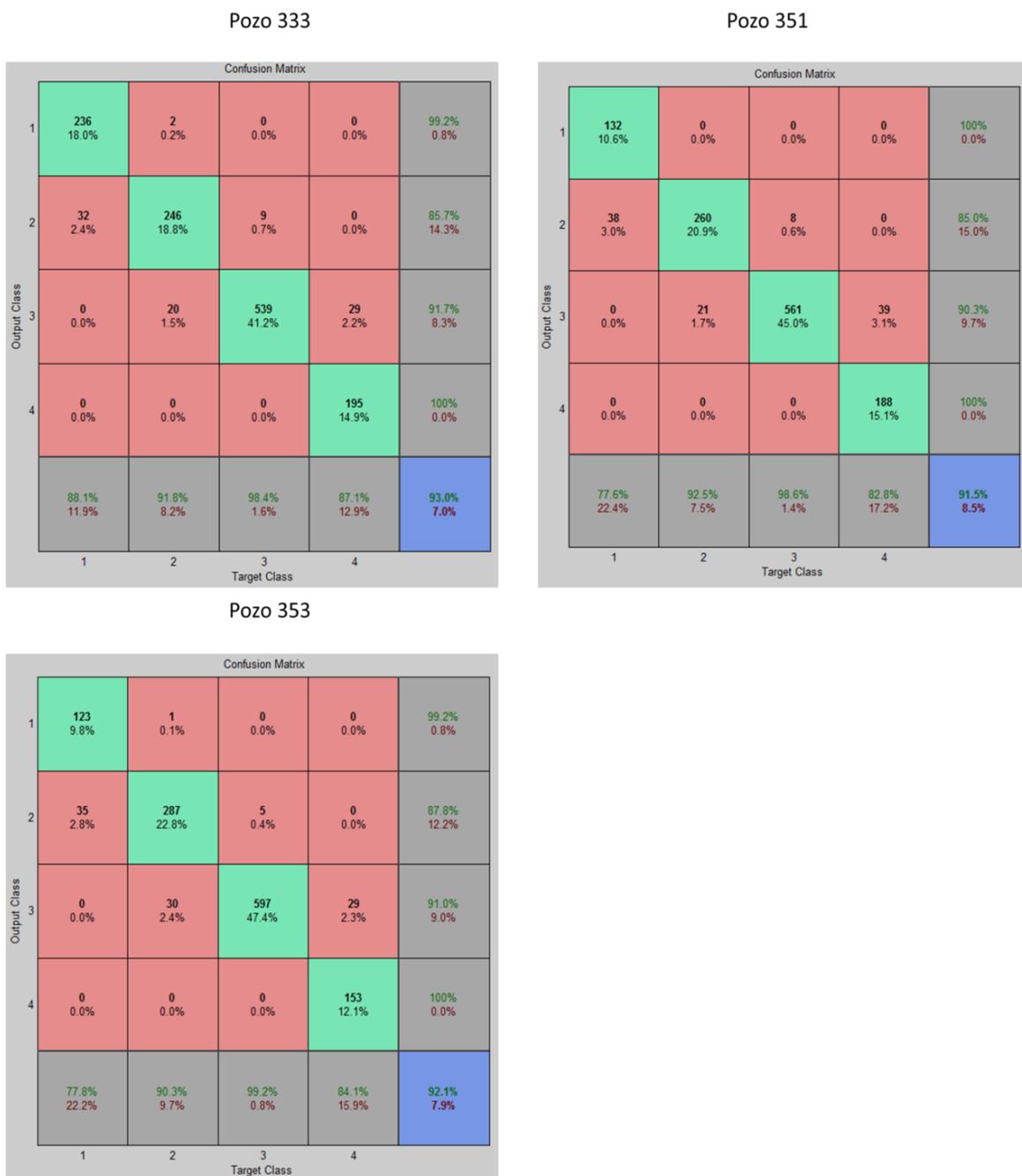


Fig. 5.20. Matrices de confusión, de la clasificación de los pozos 333, 351 y 353

Los resultados de la clasificación se muestran en las Figs. 5.21 y 5.22, con el registro de rayos gama en el carril uno, en el carril dos las litofacies estimadas por los expertos y en el carril tres (a excepción del pozo 318 que se muestra en el carril dos). Las litofacies clasificadas con la red perceptrón multicapa (MPL), se muestran bajo el siguiente código de colores:

Electrofacie - 1 Arenisca
 Electrofacie - 2 Arenisca arcillosa
 Electrofacie - 3 Lutita arenosa
 Electrofacie - 4 Lutita



Electrofacie -1 estimada con MPL
 Electrofacie -2 estimada con MPL
 Electrofacie -3 estimada con MPL
 Electrofacie -4 estimada con MPL



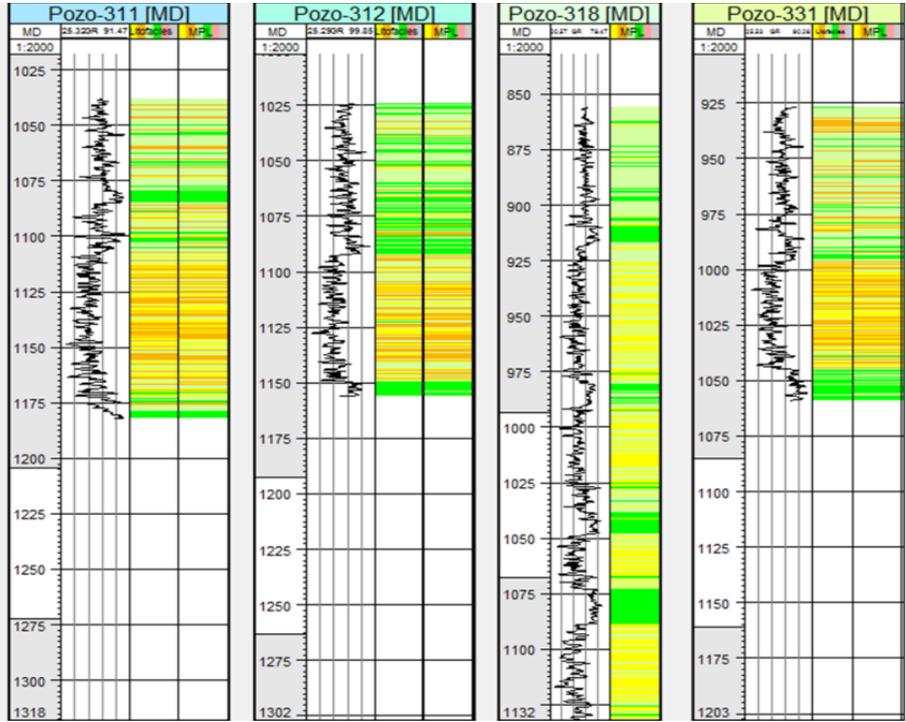


Fig. 5.21. Resultados de la clasificación usando MLP, para los pozos 311, 312, 318 y 331 (Carril 1, registro de rayos gama, carril dos la clasificación propuesta por los expertos, carril tres, clasificación mediante la red MLP)

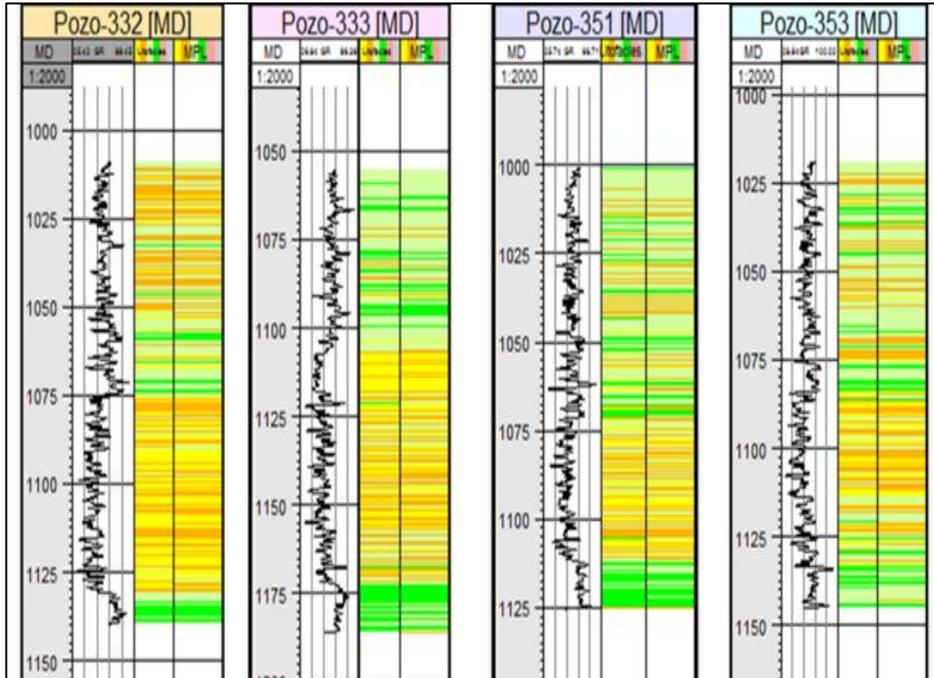


Fig. 5. 22 Resultados de la clasificación usando MLP, para los pozos 332, 333, 351 y 353 (Carril 1, registro de rayos gama, carril dos la clasificación propuesta por los expertos, carril tres, clasificación mediante la red MLP)

El grado de clasificación de los datos con este modelo neuronal es bastante aceptable (determinado por las matrices de dispersión), observando que la clasificación propuesta por expertos es bastante consistente entre los pozos.

5.4.5.2 Mapas auto-organizados.

Para el entrenamiento con los mapas auto-organizados, se procedió de igual manera que para las redes *MPL*, se entrenó la red con el pozo 333, utilizando los registros rayos gama, resistividad, caliper, sónico y neutrón. Siendo los registros de rayos gama y resistividad, los que tuvieron más peso para la obtención de la clasificación mostrada en las Figs. 5.25 y 5.26.

Los parámetros utilizados fueron:

- Un mapa de 10 x 10 neuronas (100 neuronas), en una topología hexagonal.
- Se utilizaron los datos normalizados en función de la desviación estándar y la media usando la función de Matlab “*mapstd*”.

La Fig. 5.23 muestra los pesos o distancia que tienen los datos para una determinada vecindad de neuronas.

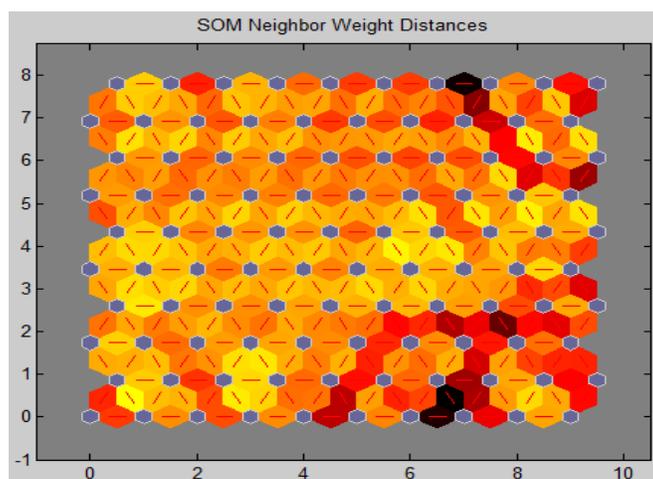


Fig. 5.23. Mapa de pesos del pozo 333, donde cada color representa el peso que le da una vecindad de neuronas a las entradas. Colores más oscuros representan menores pesos.

La Fig. 5.24 muestra los mapas de peso por variable, en ella se observa que neuronas le dan más peso a una determinada variable, donde los registros *GR*, *DT* y *NPHI*, muestran formas similares en determinadas zonas del mapa auto-organizado, siendo estas las de mayor correlación lineal.

La Fig. 5.25 muestra el mapa de éxito, donde se observan el número de datos que se encuentran en cada zona. En ella se pueden observar algunos ejemplos de grupos (clústeres) formados (la línea punteada de la Fig. 5.25)

De las matrices de confusión de las Figs. 5.26 y 5.27, los porcentajes de datos bien clasificados con respecto a las facies propuestas por los expertos es de alrededor del 50 al 80%, siendo el pozo 311, el que menor grado de reconocimiento tuvo.

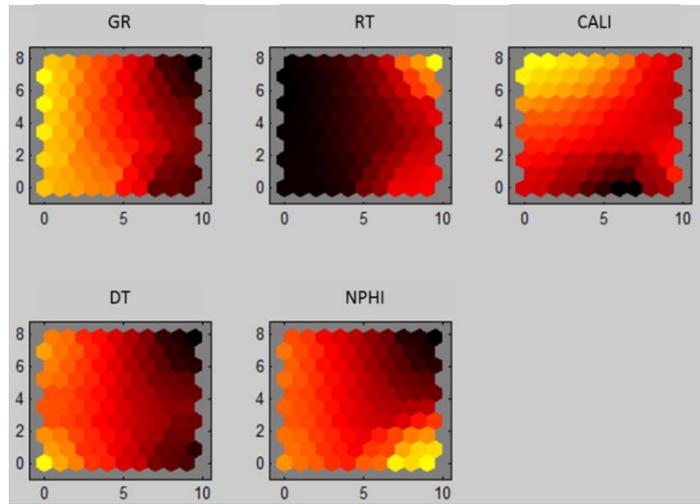


Fig. 5.24 Mapa de pesos del pozo 333, por variable, donde cada color representa el peso que le da una vecindad de neuronas a las entradas (colores más claros son mayores pesos y menores distancias)

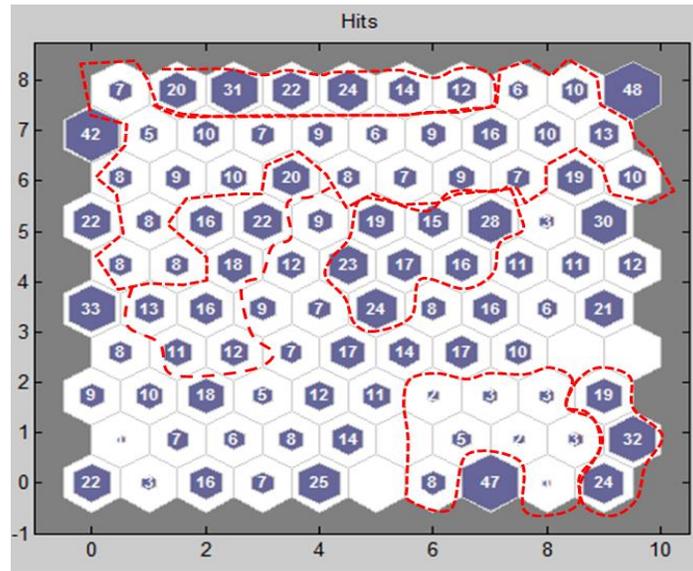
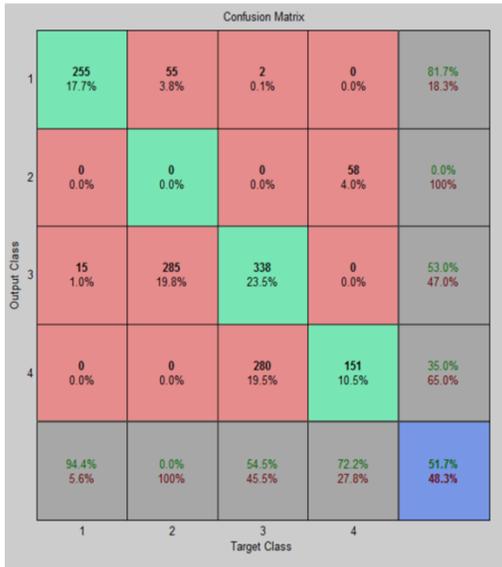


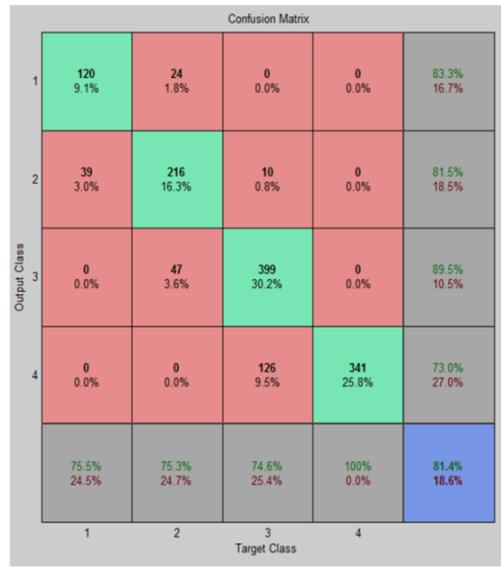
Fig. 5.25 Interpretación del mapa de éxitos del pozo 333, donde se observan algunos de los grupos (clústeres) formados en el mapa.

Los resultados de las estimaciones se muestran en las Figs. 5.27 y 5.28, en donde se puede observar que la clasificación realizada mediante los mapas auto-organizados, responde muy bien a la clasificación propuesta por los expertos del IMP, si se le da un mayor peso a los registros de rayos gama y resistividad durante la clasificación (ya que es el criterio utilizado por los especialistas para una clasificación basada en el contenido de arcilla).

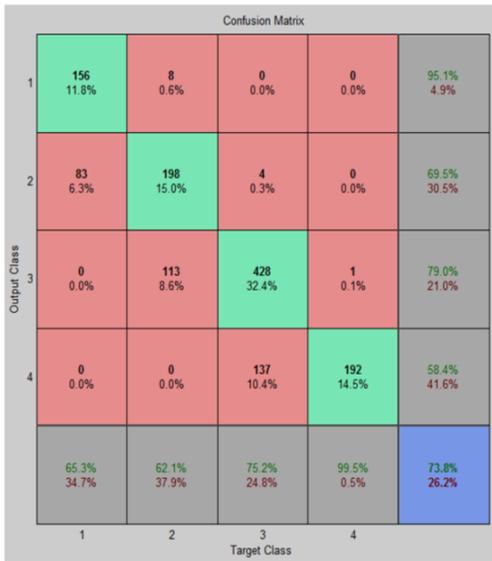
Pozo 311



Pozo 312



Pozo 331



Pozo 332

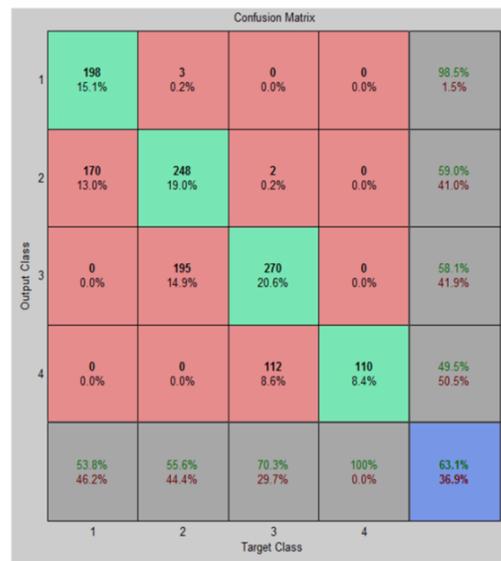


Fig. 5.26. Matrices de confusión, de la clasificación de los pozos 311, 312, 331 y 332.



Fig. 5.27. Matrices de confusión, de la clasificación de los pozos 333, 351 y 353.

Las Figs. 5.28 y 5.29 muestran en el carril uno el registro de rayos gama, en el carril dos las litofacies estimadas por los expertos y en el carril tres las litofacies resultado de la clasificación utilizando los mapas auto-organizados (a excepción del pozo 318 que se muestra en el carril dos), bajo el siguiente código de colores:

- | | | | | | |
|--------------------------------------|--|--|--|----------------------------|--|
| Electrofacies - 1 Arenisca | | | | Grupo - 2 estimada con SOM | |
| Electrofacies - 2 Arenisca arcillosa | | | | Grupo - 3 estimada con SOM | |
| Electrofacies - 3 Lutita arenosa | | | | Grupo - 4 estimada con SOM | |
| Electrofacies - 4 Lutita | | | | | |
| Grupo - 1 estimada con SOM | | | | | |

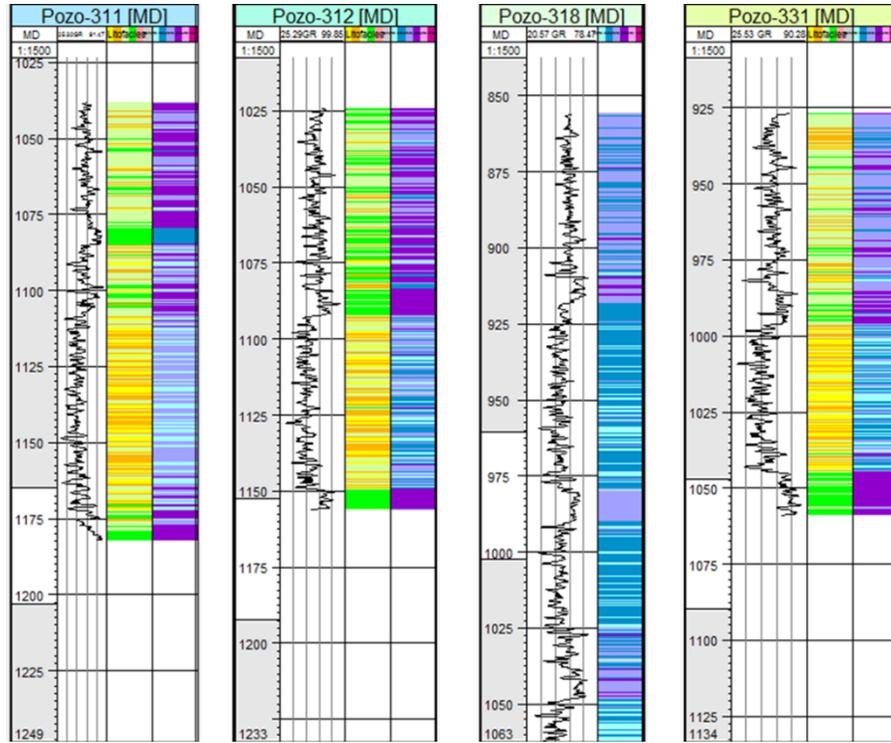


Fig. 5.28. Resultados de la clasificación usando SOM, para los pozos 311, 312, 318 y 331 (Carril 1, registro de rayos gama, carril dos la clasificación propuesta por los especialistas, carril tres, clasificación mediante la red SOM)

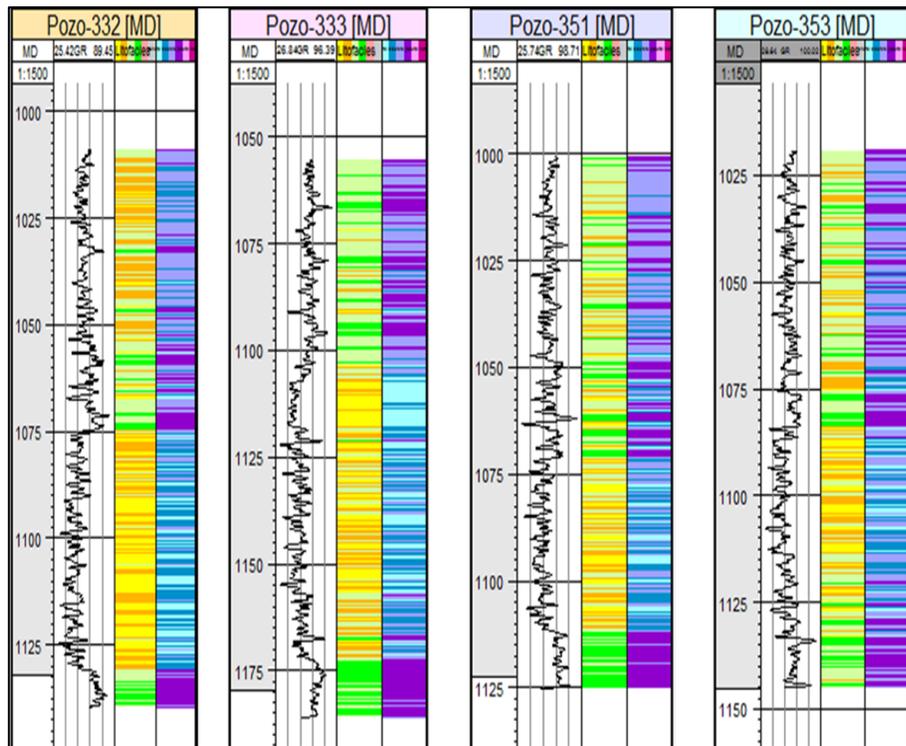


Fig. 5.29. Resultados de la clasificación usando SOM, para los pozos 333, 351 y 353 (Carril 1, registro de rayos gama, carril dos la clasificación propuesta por los especialistas, carril tres, clasificación mediante la red SOM)

5.4.5.3 Máquinas Comité.

Para el uso de las máquinas comité, se utilizaron los mismos modelos y códigos que los utilizados en la sección 4.4.5.5, de este trabajo de tesis.

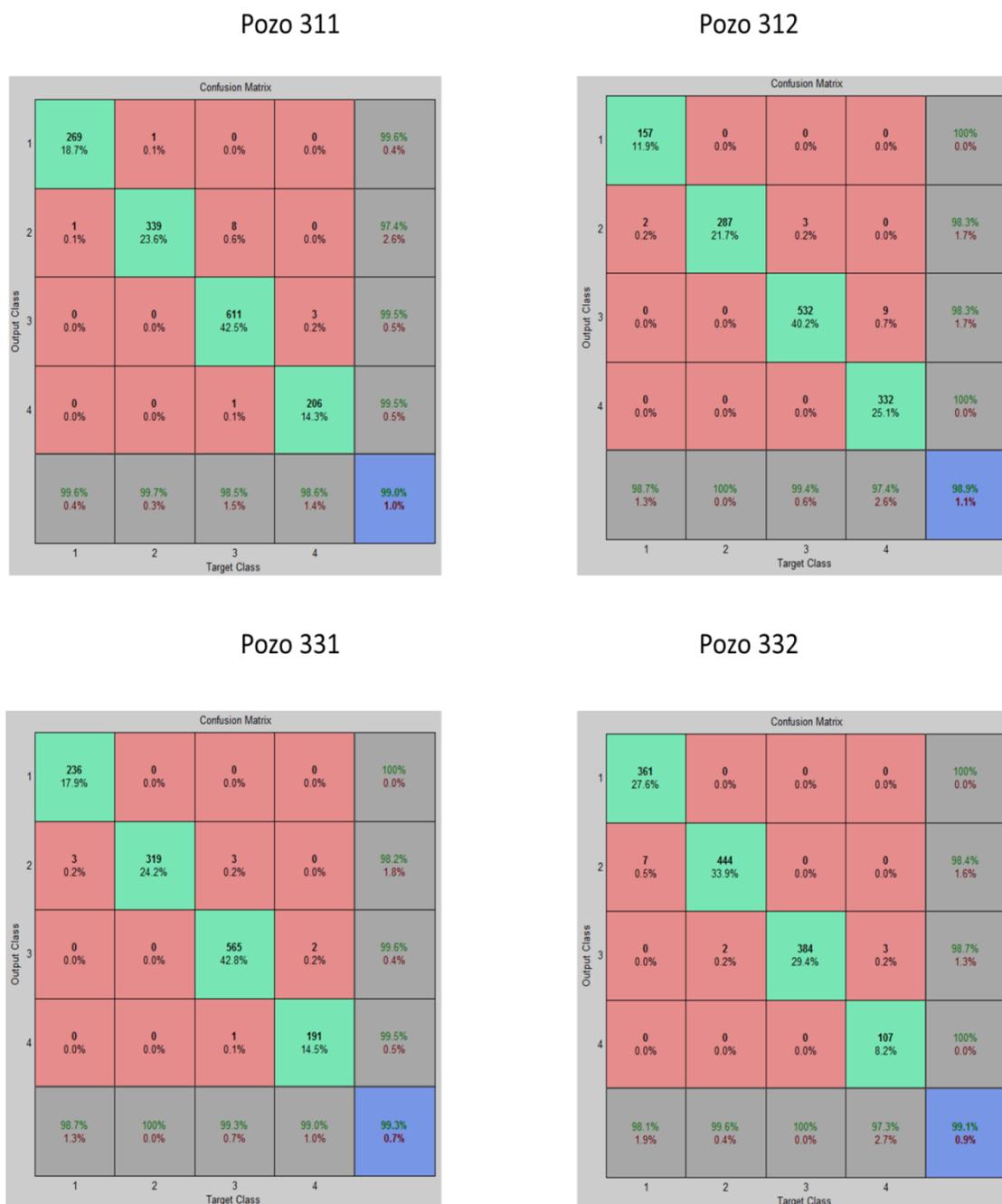


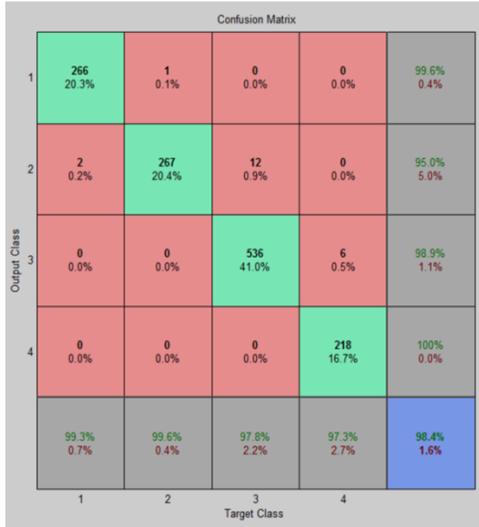
Fig. 5.30. Matrices de confusión, de la clasificación de los pozos 311, 312, 331 y 332.

Para el entrenamiento con las máquinas comité (CM), se realizaron diversas pruebas, utilizando los criterios y parámetros del análisis del modelo de perceptrón multicapa de la sección 5.4.5.1 y los

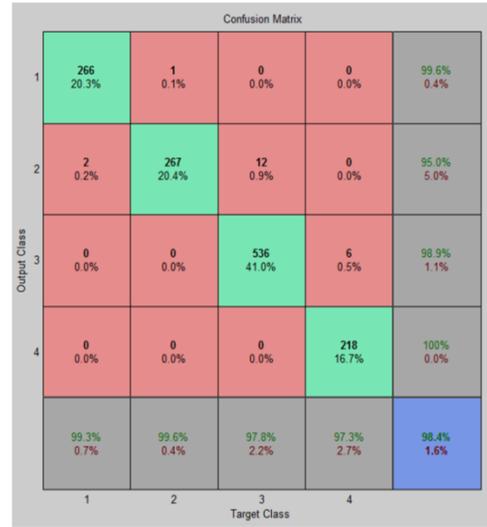
resultados del análisis de datos sintéticos. Se utilizó el pozo 333 para el entrenamiento, bajo el siguiente modelo:

La máquina comit  consta de cuatro m dulos (cuatro litofacies), cada uno de ellos formado por tres redes neuronales perceptr n multicapa simples, entrenadas con los registros y las salidas de una red neuronal de conjunto (ensemble neural network). Cada red neuronal de conjunto est  conformada por tres redes perceptr n multicapa simples, que tienen por entradas  nicamente los datos de los registros de pozo (ver Figs. 3.6, 3.7, 3.8 y 3.9).

Pozo 333



Pozo 351



Pozo 353

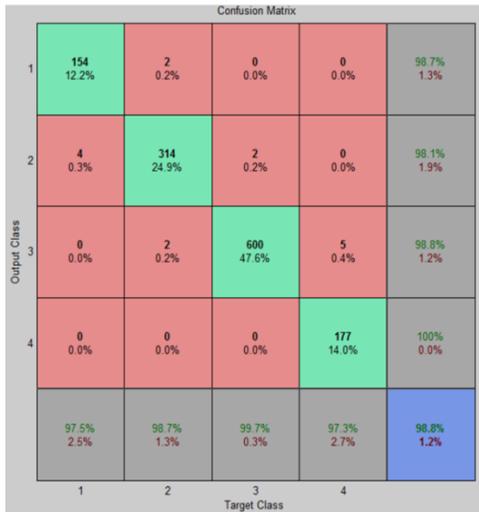


Fig. 5.31. Matrices de confusi n, de la clasificaci n de los pozos 311, 312, 331 y 332.

La diferencia entre las diferentes unidades individuales (MLP) que conforman la red modular o m quina comit  (CM), fueron el n mero de neuronas para un mismo n mero de capas, los algoritmos de

entrenamiento, la forma en que los datos son presentados a la red y los parámetros propios de los algoritmos de entrenamiento.

Las matrices de confusión de las Figs. 5.30 y 5.31, muestran los porcentajes de datos bien clasificados con respecto a las facies propuestas por los especialistas, con un rango alrededor de 98 al 99 %, siendo el pozo 333 y 353, los que menor grado de reconocimiento tuvieron.

Los resultados de las estimaciones se muestran en las Figs. 5.32 y 5.33, en donde se observa el registro de rayos gama en el carril uno, las litofacies estimadas por los expertos en el carril dos y en el carril tres las litofacies resultado de la clasificación utilizando las máquinas comité (a excepción del pozo 318 que se muestra en el carril dos). Las litofacies estimadas se muestran bajo el siguiente código de colores:

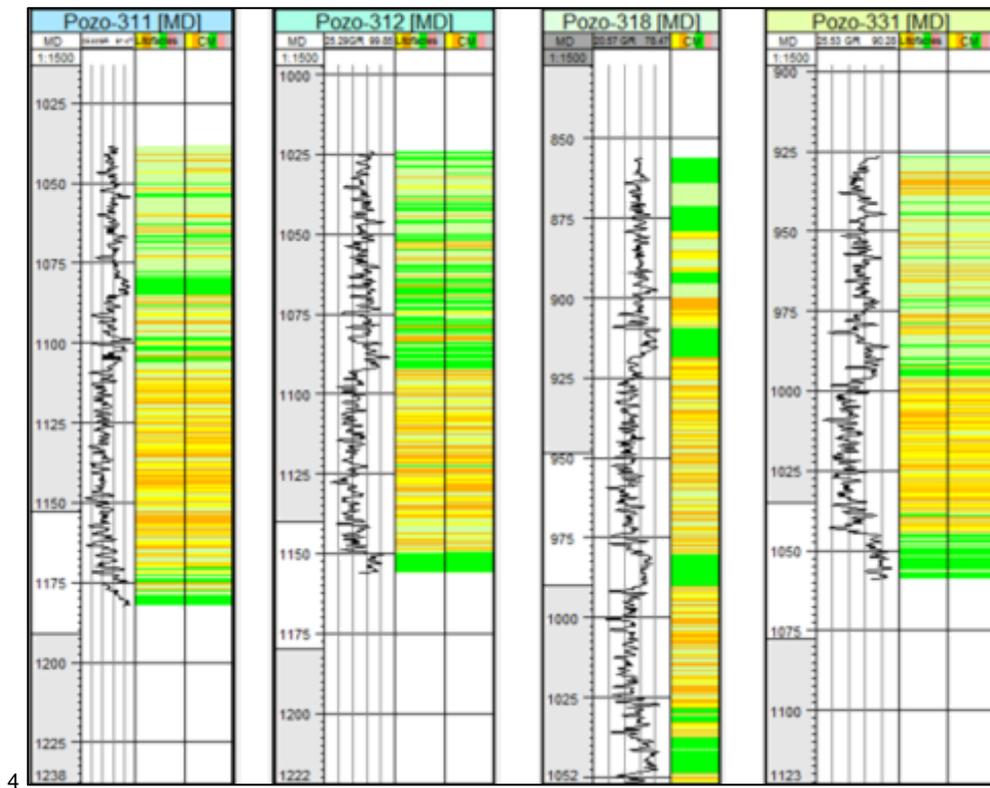


Fig. 5.32. Resultados de la clasificación usando CM, para los pozos 311, 312, 331 y 3.32. (Carril 1, registro de rayos gama, carril dos la clasificación propuesta por los especialistas, carril tres, clasificación mediante la red CM)

En las Figs. 5.32 y 5.33, se puede observar que la clasificación realizada mediante las máquinas comité, responde muy bien a la clasificación propuesta por los especialistas.

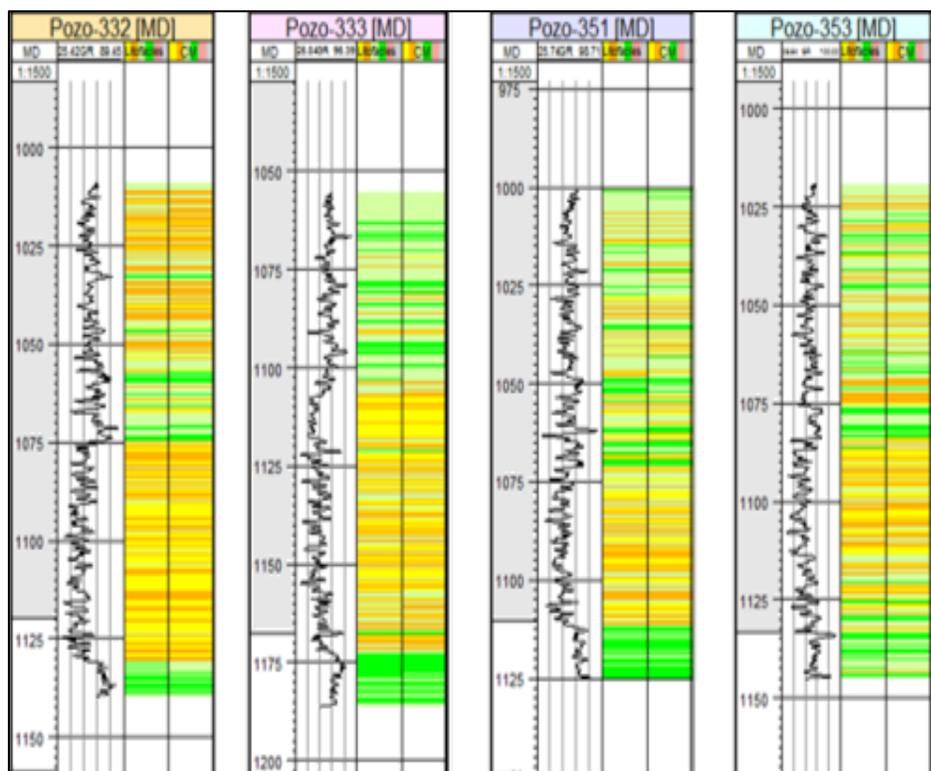


Fig. 5.33. Resultados de la clasificación usando *CM*, para los pozos 333, 351 y 353 (Carril 1, registro de rayos gama, carril dos la clasificación propuesta por los especialistas, carril tres, clasificación mediante la red *CM*)

5.4.5.4 Conclusiones del entrenamiento con redes neuronales.

En el caso de los registros reales, los tres modelos neuronales propuestos tienen consistencia con las interpretaciones realizadas por los especialistas del IMP, que conjuntaron criterios petrofísicos, y geológicos (lito y crono-estratigráficos), las cuales honran los datos. El grado de clasificación obtenido con las redes neuronales en relación con la clasificación propuesta por los expertos, puede ser catalogado de regular a muy bueno debido a que los especialistas utilizaron como principal criterio el contenido de arcilla, siendo el registro de rayos gama la variable principal que aporta elementos a la red como criterio de clasificación (conclusión a la que se llega del análisis exploratorio de datos, análisis de agrupamiento (clústeres) y análisis de componentes principales).

Como se ha mencionado a lo largo de este capítulo, los registros de rayos gama, de resistividad, caliper, sónico y neutrón, son los registros que mayor información aportan, al momento de definir una clasificación basada en el contenido de arcilla. De manera global todos los pozos son muy similares y respondieron de manera muy similar a los modelos neuronales (a excepción del 318, que presentan características algunas características diferentes a los demás pozos y el pozo 311, que carecía de todo el conjunto de registros). Las máquinas comité representaron el mayor grado de clasificación (matrices de confusión), debido a la mejora que representan en el rendimiento de las redes (el uso redundante de datos y la descomposición en subtarear, es clave dentro del proceso de optimización del entrenamiento) y la naturaleza de los datos (ya que una simple red neuronal como las perceptrón multicapa los resuelve con un grado de aceptable de clasificación). Los mapas auto-organizados tuvieron mayor dificultad para clasificar los datos, sin embargo, se encuentran en un el rango es aceptable de clasificación

Finalmente se puede concluir que el conocimiento de los datos mediante los análisis realizados, permitió conocer las variables de entrada que mejor clasificación generarían, también nos permitió validar los resultados con un grado de confiabilidad alto y entender los fenómenos asociados a la información, así

como también interpretar los resultados y justificar lo que exactamente está pasando, durante el uso de las redes neuronales como clasificadores de datos.

5.4 Interpretación y comentarios de los resultados con redes neuronales

En general se pueden identificar los intervalos de dos secuencias, con sus respectivos cuerpos de arcillosos (Fig. 5.34), estas secuencias se componen principalmente de arenas asociadas a lóbulos en facies de abanico medio o distal (facies G y H de la Fig. 4.1). La secuencia areno-arcillosa, está conformada por arenas y arenas arcillosas intercaladas con lutita, como parte de secuencias turbidítica. Los diferentes eventos se interpretan como resultado del descenso del nivel medio del mar (durante la formación del abanico de piso de cuenca). La Fig. 5.34 muestra la sección de pozos con las diferentes clasificaciones. En esta Fig. se muestra en el carril uno el registro de rayos gama, en el carril dos la clasificación realizada por los especialistas del IMP, en el carril tres la clasificación mediante redes perceptrón multicapa, en el carril cuatro la clasificación mediante mapas auto-organizados y finalmente en el carril cinco la clasificación mediante el uso de las máquinas comité.

Una parte importante de los análisis es buscar la correlación con la información sísmica mediante el amarre del pozo con la sísmica y la información de núcleos, con el fin de identificar con mayor precisión, el tipo de facies de depósito, mediante la correlación cronoestratigráfica y su relación con los diferentes litotipos o litofacies. Para así definir un modelo de facies consistente (a escala de pozo y a escala sísmica), para el modelado geológico petrofísico.

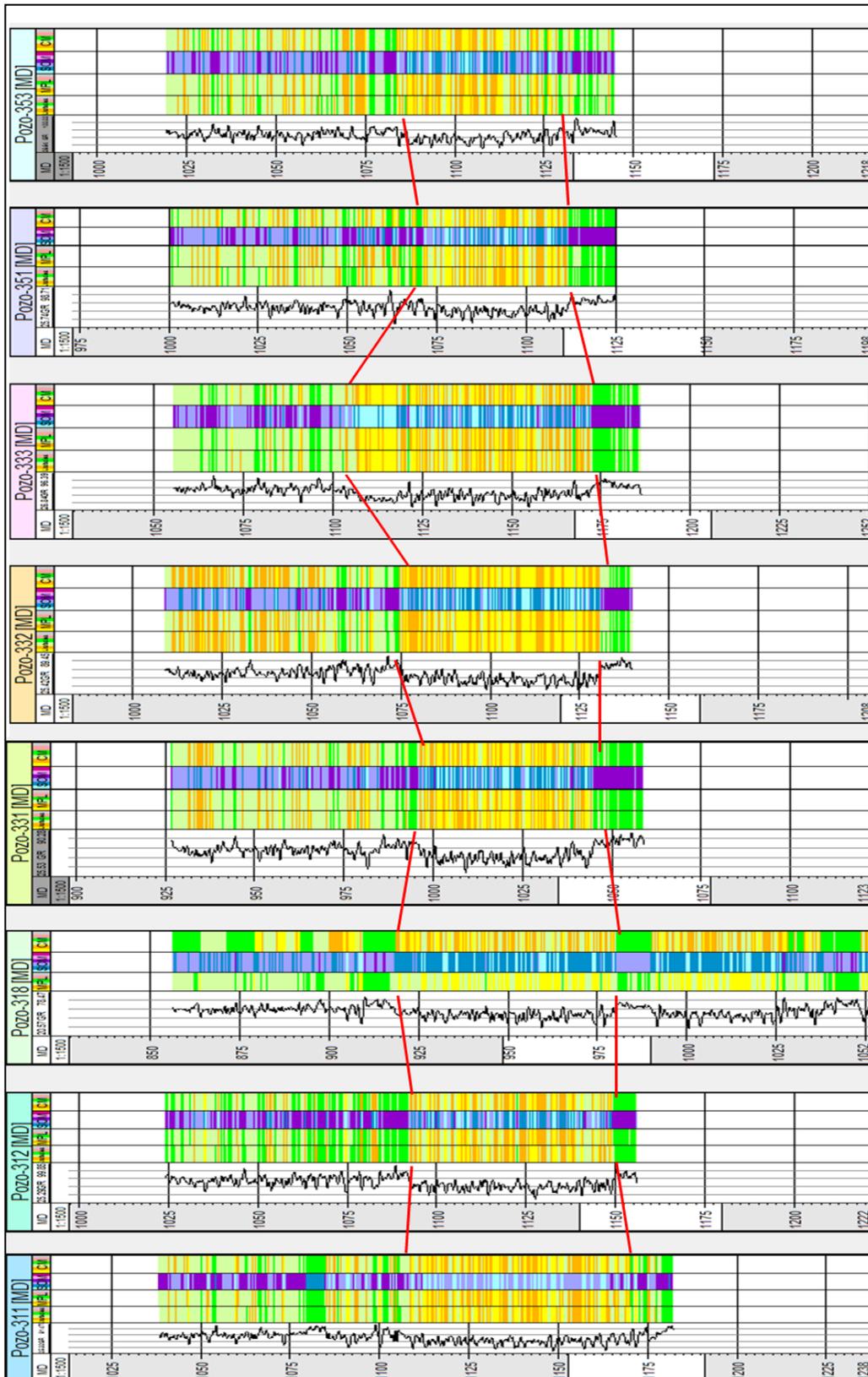


Fig. 5.34. Sección de correlación de secuencias utilizando la clasificación de redes neuronales (carril uno rayos gama, carril dos litofacies originales, carril tres litofacies-MLP, carril cuatro litofacies-SOM y carril 5 litofacies -CM)

VI. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

A continuación se presenta el análisis de los resultados obtenidos de la metodología planteada en este trabajo, aplicada a los datos sintéticos y al caso de estudio.

Los datos de entrada utilizados en este estudio, corresponden a un conjunto de registros convencionales de pozos (Rayos gama, resistividad, caliper, densidad y neutrón), los cuales son el conjunto más común para realizar evaluaciones petrofísicas básicas, definir probables litologías y establecer sistemas de depósito (mediante patrones de las curvas rayos gama y resistividad). Para este trabajo, este conjunto de registros fue analizado mediante el análisis exploratorio de datos, el análisis de grupos y análisis de componentes principales, con el objeto de realizar un control de calidad de los datos, entender el comportamiento estadísticos de los mismos en función de la clasificación propuesta de los datos.

De los análisis se observó que los registros sintéticos y los registros reales, en general presentan condiciones diferentes, ya que es muy difícil modelar registros geofísicos que representen la realidad física de registros en un campo tan complejo como lo son los campos del Paleocanal de Chicontepec. Hay que recordar que los registros sintéticos fueron generados a partir de las expresiones propuestas en el capítulo cuatro y que fueron utilizados los valores medios de registros de pozo de cuerpos arenosos en el campo Spraberry (Texas, E.U.); registros que muy probablemente fueron tomados con condiciones completamente diferentes y/o editados con diferentes criterios. No obstante lo anterior, se puede comprobar algo que a primera instancia parecería lógico, sin embargo, muchas veces no es del todo entendido, que la clasificación de datos está en función de la naturaleza estadística de los mismos. Si para una determinada electrofacie, la varianza de una variable específica es grande, implicará que su rango de valores sea amplio, lo cual provocará que los límites entre electrofacies estén muy cercanos o que estén superpuestos, lo cual generará confusión a cualquier estimador al que le sean presentados valores de ésta variable, cercanos a los límites de dicha electrofacie; ya que podrán parecerse a valores que pertenecen a otra electrofacie. Es por ello que cada variable tiene grado de mayor a menor importancia, para definir las electrofacies y el objetivo del análisis estadístico es reconocer esta importancia.

El análisis de los datos sintéticos planteó tres problemas específicos: datos bien clasificados, medianamente clasificados y mal clasificados, en función de los rangos de los datos y sus varianzas, tratando de emular el comportamiento real de estudio. Un factor clave que juega en los datos sintéticos es el efecto del ruido aleatorio para degenerar la clasificación de los datos, en función de sus valores.

Del análisis estadístico de los registros sintéticos, se observó que el modelo tres es estadísticamente diferente a los modelos uno y dos, los cuales eran relativamente más parecidos. Es por ello que los sistemas no supervisados (k-medias y SOM) tienen una relativa ventaja sobre los métodos supervisados (MPL y CM), al tratar de clasificar el modelo tres, a diferencia de los modelos uno y dos que son más parecidos y donde los métodos supervisados tuvieron mejores resultados en la clasificación, ya que se requiere de un maestro o guía para clasificar y mientras este sea diferente de lo que se desea clasificar, se obtienen resultados erróneos. Prueba de lo anterior son los tres ejercicios realizados con las redes perceptrón multicapa (MPL): a) entrenar con el modelo uno y clasificar los demás modelos, b) entrenar con el modelo dos y clasificar los demás modelos y finalmente c) entrenar con el modelo tres y clasificar los demás modelos; donde claramente se observa que los modelos uno y dos tienen un muy buen relativo grado de clasificación, entrenando con uno u otro modelo, pero fallan al clasificar el modelo tres. Lo mismo sucede al entrenar con el modelo tres, en el cuál, resulta prácticamente imposible clasificar correctamente los modelos uno y dos.

Al analizar los datos de los pozos del caso en estudio, se observó en el comportamiento del pozo 318, donde las características de los registros de pozo parecen ser distintas a las de los demás pozos (análisis por pozo, el cual no es mostrado). Es por ello que los expertos, no realizaron una clasificación similar en este pozo, a la usada en los demás, utilizaron un criterio más petrofísico. Sin embargo se logra observar un grado de clasificación coherente con los demás pozos.

Es necesario aclarar que en el análisis de los registros sintéticos, todos los métodos de clasificación, confunden la electrofacie tres debido a que el criterio de clasificación es más asociado al patrón geométrico del registro de rayos gama y al ser una secuencia de arenas con lutitas, tienden a ser discretizada en secuencias de la electrofacie uno (arenas) y electrofacie dos (lutitas). Es importante señalar también que las malas clasificaciones están asociadas principalmente a que las arenas de los diferentes modelos presentan condiciones petrofísicas diferentes (diferente porosidad, volumen de arcilla, etc.) y diferente contenido de fluido, más sin embargo todos los sistemas de clasificación logran diferenciar los diferentes cuerpos (secuencias), lo mismo sucede en el caso de los registros reales, en donde es posible definir diferentes secuencias de depósito, desde un punto de vista más regional (es decir a un nivel de escala más global).

En los registros sintéticos la electrofacie dos (lutitas) es la que menor problema tuvo para su clasificación, ya que las características de esta son iguales para los tres modelos, lo cual resulta importante ya que los modelos neuronales responden principalmente en facies arcillosas, mismo que se comprueba con los resultados mostrados en los pozos del caso en estudio.

En el caso de los registros sintéticos las variables que sirvieron a una mejor clasificación fueron la relación RT/GR (principalmente por que el criterio para definir las diferentes secuencias fueron los registros de rayos gama y resistividad), el registro caliper (debido a que existe diferencia entre las electrofacies arcillosas y las no arcillosas), el registro sísmico (que responde principalmente a la matriz y las características de la matriz son prácticamente iguales en los tres modelos) y el registro de neutrones que si responde más a los fluidos y permite diferenciar entre un tipo dado de arenas. El registro de densidad prácticamente fue descartado, debido a que no mostraba un gran contraste de valores entre electrofacies (al ser siliciclásticos). La curva dGR/dZ , muestra los cambios de pendiente de la curva, sin embargo, al tener ruido de alta frecuencia los cambios son constantes y no es posible definir un patrón específico. Los registros RT/GR , Caliper, sísmico y neutrón, son los que menor coeficiente de correlación lineal de Pearson mostraron entre ellos y del análisis de componentes principales, permite dar una idea de la aportación a cada componente y cada componente a la varianza, con el objetivo de no redundar la información, ya que este es el aspecto clave durante cualquier proceso de clasificación. Bajo los mismos criterios se estableció en el caso de los pozos del caso en estudio que los registros a utilizar son los registros rayos gama, resistividad, caliper, neutrón y sísmico. La relación RT/GR , no se utiliza debido a que no mostro resultados óptimos durante las pruebas de entrenamiento.

Del análisis exploratorio de datos y las diferentes pruebas realizadas con redes neuronales, se puede observar que no fue necesario realizar transformaciones logarítmicas o eliminar valores atípicos para mejor la simetría de la distribución de los datos (y no alterar los datos), ya que esto no mejoraba el rendimiento de la red y la pérdida de información alteraba los resultados. No obstante fue necesario realizar la normalización de los datos para mejorar el rendimiento de la red. Se utilizó la función $mapstd$ (Ec. 3.18), la cual normaliza en función de la desviación estándar y la media, sin embargo también es posible utilizar otro tipo de normalización, por ejemplo entre cero y uno.

$$x_{norm} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad 6.1$$

Para el análisis de grupos se observaron algunos puntos importantes:

- Del análisis de dendrogramas, se definieron la existencia de subelectrofacies, existentes tanto en los datos sintéticos, como en los datos reales, lo cual se observó durante el entrenamiento con los mapas auto-organizados.
- La importancia de analizar los centroides, permite observar la similitud de las variables en términos de las distancias y por ello es necesario utilizar las variables con menor similitud (grado de correlación) entre ellas para obtener mejor grado de clasificación. Los datos sintéticos mostraban mayor diferencia en sus centroides a diferencia de los datos reales, por lo que definir un criterio o darle un peso a determinado registros es importante para obtener mejores

clasificaciones (En el caso de los registros sintéticos fue la relación RT/GR y en los datos reales el registro de rayos gama, el cual fue la base de la interpretación manual hecha por los especialistas).

- El definir el número de electrofacies es importante, para datos desconocidos, el iniciar con números reducidos de electrofacies es importante así como realizar pruebas con diferentes números de grupos para tomar una decisión.

De todos los análisis estadísticos se podría englobar el caso de estudio dentro de las características del modelo uno y dos, donde los pozos son muy parecidos entre sí, pero la diferencia entre los diferentes cuerpos en términos de los registros de pozos, están medianamente bien clasificados.

Para la parte del entrenamiento y clasificación con los modelos neuronales propuestos, los parámetros que más influyeron en la obtención de una buena clasificación en términos de la clasificación manual, fueron en orden de importancia: Los datos de entrada (es importante utilizar los registros que estadísticamente sean los más adecuados), el pre-procesamiento de los datos (ya que las redes neuronales, si los datos no están normalizados, tienden a arrojar resultados sin sentido) los conjuntos de datos de entrenamiento, validación y prueba (en el caso de los modelos no supervisados), número de capas ocultas y neuronas (del análisis de registros sintéticos, dos capas ocultas, con diez y cuatro neuronas respectivamente) y finalmente los parámetros de entrenamiento, los cuales principalmente se establecieron de las pruebas de ensayo y error.

Analizar los datos directamente de Matlab no fue tarea sencilla, sin embargo el análisis exploratorio de datos fue crucial para dar el peso a la variable que permitiría establecer el criterio para dar una clasificación final. Dicha clasificación fue reinterpretada de manera conceptual, para poder comparar de manera gráfica los resultados con la clasificación manual, por lo que el criterio del intérprete es crucial, para generar el producto final que servirá de insumo durante el modelo de litofacies en la Modelación Geológica Petrofísica de Yacimientos.

En el caso de estudio los tres modelos neuronales propuestos reproducen de manera muy buena las interpretaciones manuales, pero son las máquinas comité, las que menor grado de error presentaron al analizar las matrices de confusión (dando mayor peso al registro de rayos gama el de mayor peso), a diferencia de los registros sintéticos donde estos modelos (MPL y CM) clasifican únicamente los modelos uno y dos. Las Maquinas comité explotan las propiedades que tienen las redes de generalizar y memorizar (sobre-entrenar), además de facilitar la tarea mediante la subdivisión de tareas, es quizá por ello que el rendimiento fue mejor en el caso de los datos reales.

De los resultados con las redes neuronales se pudo analizar la consistencia de las clasificaciones originales (los criterios de clasificación van implícitos), por ejemplo, la electrofacie tres de los modelos sintéticos, quizá no deba ser una electrofacie diferente, sino una sucesión de electrofacie uno y dos (arenas y lutitas), En el caso de los pozos parece que existe congruencia con la clasificación manual en todos los pozos, si el pozo 318 fuese clasificado manualmente utilizando los mismos criterios, es posible que no se hubiera podido reproducir con los modelos neuronales.

Los resultados de cualquier proceso automatizado deben ser analizados e interpretados por analistas, para ser integrados con toda la información disponible, para así generar un modelo de litofacies congruente con el modelo geológico conceptual de un área de estudio, si esto no es posible, es probable que los criterios de clasificación no sean los más adecuados y es necesario regresar a definirlos. Los resultados de todos los análisis (los análisis estadísticos y clasificación redes neuronales), pueden dar pauta a establecer dichos criterios. La integración y la congruencia de los resultados, con otra información quedan fuera del alcance de los objetivos de este trabajo de tesis, sin embargo es una parte crucial, ya que lo importante dentro de cualquier estudio que necesite un modelo de facies válido, es el producto final.

VII. PROPUESTA METODOLÓGICA.

Hasta el momento se han descrito cada uno de los pasos que conforman la metodología de trabajo y se mostraron los resultados al aplicarla a dos casos de estudio uno de datos sintéticos y otro de datos reales. A partir de los resultados obtenidos en este trabajo, se puede establecer una propuesta metodológica, fundamentada en criterios establecidos a lo largo de este trabajo y en las condiciones que los mismos datos plantean.

7.1 Establecer un marco geológico de referencia.

Lo más importante de la metodología es tomar en cuenta el marco geológico de referencia, a través de la comprensión de los modelos conceptuales establecidos por las diferentes disciplinas que integran un grupo de trabajo multidisciplinario, así como el entendimiento de proceso de integración de los mismos, a partir de la revisión y estudio de la información disponible en un área determinada.

7.2 Establecer los criterios de clasificación.

Independientemente del modelo de red neuronal a utilizar, sea éste supervisado o no supervisado, es sumamente importante contar con una clasificación manual (en el caso supervisado servirán de guía o patrón de entrenamiento y en el caso no supervisado, son puntos de calibración) realizada por profesionales expertos y validada por un grupo de trabajo multidisciplinario, que sea integral y coherente con los modelos conceptuales realizados por dicho grupo, realizada en los pozos más representativos del área (el análisis exploratorio de datos es un herramienta poderosa para determinarlo) y con información de núcleos como referencia. Es importante conocer además, cuáles fueron los criterios para realizar dicha clasificación y entender, al menos de manera general, el proceso de clasificación manual y el esquema de acondicionamiento de las curvas de los registros realizadas por un petrofísico (si completó el registro, si se realizó suavizado de las curvas, si realizó correcciones ambientales, si normalizó los histogramas, etc.)

7.3 Control de calidad de los datos.

Independientemente del proceso realizado por el petrofísico, es importante revisar los archivos que son los insumos del proceso, revisando si no están corruptos, si las curvas están completas, si hay picos, si el nombre de la variable (acrónimo del registro) en realidad representa el registro que se quiere etc. Una forma de hacerlo es graficarlos.

También es importante organizar la información, revisar con que curvas se cuenta y cuáles no, así como identificar si los acrónimos para un determinado tipo de registro son equivalentes en todos los pozos, es decir, si se cuenta con curvas equivalentes en los pozos para una determinada propiedad, independientemente de la herramienta con que se haya medido.

7.4 Análisis exploratorio de datos

Se recomienda realizar un primer análisis exploratorio de datos realizado de manera conjunta con geólogos y petrofísicos expertos durante la etapa de clasificación manual, con la información de registros de todos los pozos y por pozo de manera individual. Lo anterior permite definir los criterios de clasificación y entender el comportamiento general de los datos, así como una herramienta de control de calidad del dato mismo.

Se propone realizar un segundo análisis exploratorio de datos, realizado por clase (electrofacies o cualquier criterio de clasificación de registros) y por pozo-clase, con el objetivo de validar la clasificación y obtener información de la misma.

Se propone además un tercer análisis exploratorio de datos, nuevamente realizado por clase y por pozo – clase, dependiendo del segundo análisis exploratorio, si existe la necesidad de transformarlos o eliminar

valores atípicos. Cuando no es necesario realizar algún tipo de transformación de tipo logarítmico, se recomienda utilizar datos normalizados durante el entrenamiento de una red neuronal (normalización entre cero y uno o una normalización basada en la media y la desviación estándar).

Para los análisis exploratorios de datos se propone:

- a) Analizar y revisar los de los estadígrafos calculados para cada variable, en forma de tablas.
- b) Analizar histogramas y gráfico de caja, para identificar simetrías y posible comportamiento multimodal.
- c) Analizar gráficos de caja en un solo gráfico, para comparar rangos intercuartílicos, diferencias en la media y mediana, así como la presencia de datos atípicos.
- d) Analizar gráficos cuantíl – cuantíl, para analizar la posibilidad de un comportamiento normal.
- e) Realizar los gráficos de dispersión y analizarlos en forma de matriz, junto con sus coeficientes de correlación lineal de Pearson (puede probarse algún otro criterio de correlación), para identificar posibles grupos (clústeres) y conocer el panorama general de clasificación que proporciona las variables, de manera bivariada.

No es necesario realizar un análisis de regresión, ya que los modelos neuronales suelen ser interpretados como procesos de regresión multivariada.

7.5 Análisis de agrupamiento.

Después del análisis exploratorio de datos, se propone realizar un análisis de dendrogramas, con la finalidad de analizar los posibles grupos a formarse y tener una idea clara de la representatividad de los datos en cada uno de esos grupos. Es importante analizar detenidamente el comportamiento de los dendrogramas en función de los criterios utilizados durante la clasificación manual y se propone comparar con el uso de algún análisis de agrupamiento jerárquico, en los cuales no se conoce a priori el número de grupos a formar (el cual no fue utilizado en este trabajo debido a que el número de grupos suele ser alto).

Ya que el objetivo, es validar la clasificación manual se propone principalmente utilizar un análisis no jerárquico (en este caso fue k-medias), en los cuales se puedan definir el número de grupos a formar y con ello tener idea de la consistencia de la clasificación original. Se pueden utilizar algunos criterios para definir el número de grupos (criterio de Información de Akaike, criterio de Inferencia Bayesiana, etc.) y comparar con la clasificación original.

Una vez elegido el número de grupos y un criterio de similitud, se realiza una clasificación rápida por el método de las k medias, con el objetivo de comparar con la clasificación manual y obtener los centroides de los grupos y el centroide global. El análisis de los centroides mediante gráficos de dispersión (los cuales se proponen analizar, analizando por color cada uno de los puntos clasificados por el método de k medias), ayuda a entender de manera cuantitativa, las variables o pares de ellas que permiten tener mejor separación en función de la distancia y con ello comenzar a dar idea de los problemas que enfrentará la red neuronal al momento de elegir una u otra variable. También se recomienda analizar de individual de cada variable los gráficos de punto, mostrando en ellos los puntos que pertenecen a uno u otro clúster.

Los productos finales de esta etapa son:

- a) El número de grupos que se esperaría fuesen reproducidos por el modelo neuronal.

- b) La clasificación validada o desechada en función de la distancia (así como una interpretación de lo anterior)
- c) Una idea cada vez más clara de las variables que contribuirán a una mejor clasificación.

7.6 Análisis de componentes principales

Para el análisis de componentes principales se propone el análisis de los estadísticos de correlación de las variables, así como de la contribución de las variables a las componentes y la contribución de dichas componentes a la varianza. Se propone también analizar los gráficos de sedimentación y se deja de manera opcional el uso de las componentes como variables preprocesadas que serán la entrada de las redes neuronales (esto en función de los ensayos realizados con los modelos neuronales).

El análisis de biplots puede ser opcional y sin embargo se recomiendan ya que son gráficos útiles que resumen e integran los análisis previos (análisis de varianza, de correlación, análisis de agrupamiento y de componentes principales).

7.7 Aplicación de las redes neuronales.

Es importante reafirmar que después de los análisis mostrados en este trabajo de tesis, el punto más importante para el buen funcionamiento de los sistemas neuronales y la obtención de resultados aceptables, son los datos de entrada. Sin importar el tipo red a usar, su topología, la división de los conjuntos de datos para entrenar, validar o probar, los parámetros de los algoritmos de entrenamiento, el principal parámetro para un buen funcionamiento es el uso de las variables adecuadas y el preprocesamiento que se les pueda realizar.

En este contexto se propone para el entrenamiento:

- a) Seleccionar las variables en función de los análisis estadísticos previos, usualmente los registros utilizados comúnmente para discriminar litologías
- b) Normalizar los datos, ya sea entre cero y uno o utilizando algún criterio, como el empleado por MATLAB (función `mapstd`), en donde se realiza la normalización en función de la desviación estándar y la media. El criterio de utilizar uno u otro, está en función de los datos y las pruebas realizadas.
- c) Dividir el conjunto de entrenamiento en función del modelo neuronal a utilizar. En el caso de las redes perceptrón multicapa y los mapas auto-organizados, la intención es evitar la redundancia y mejorar la generalización. En el caso de las máquinas comité, la intención es redundar y sobreentrenar con las redes perceptrón multicapa en un esquema de redes de conjunto y posteriormente evitarla usando el esquema modular.
- d) El número de capas ocultas y neuronas está en función de los resultados del uso de datos sintéticos y de no disponerse, se pueden realizar pruebas tomando en cuenta los objetivos de sobreentrenar la red (mayor número de neuronas) o mejorar su generalización (menor número de neuronas).
- e) Desgraciadamente para el uso de los diferentes algoritmos de entrenamiento y sus diferentes parámetros, es necesario realizar diferentes pruebas.
- f) Es necesario comprobar los resultados de la clasificación, si se cuenta con puntos de calibración o amarre (en el caso supervisado de las guías o patrones de entrenamiento), utilizando las matrices de confusión, las cuales son herramientas para calificar el grado de clasificación de nuestra red neuronal.

- g) Realizar una comparación visual y secciones de pozo con las litofacies clasificadas, para corroborar la congruencia de los resultados.

En este trabajo se recomienda calificar los datos en función de los sistemas de bien a mal clasificados, con la idea de clasificar primero los pozos que estadísticamente sean más parecidos y que tengan patrones o guías de entrenamiento; con la idea de definir el modelo neuronal más óptimo.

En el contexto anterior se recomienda:

- a) El uso y entrenamiento de mapas auto-organizados en datos que no estén muy bien clasificados, que no tengan patrón o guía de entrenamiento y que sean estadísticamente muy diferentes a los demás pozos que se tengan controlados. La mejor forma de entrenarlos es usar directamente los datos del mismo pozo y clasificarlo, aunque se pueden clasificar usando un modelo de mapas auto-organizados entrenado en otro pozo, con información de núcleos y/o clasificación manual de amarre.
- b) Para los demás pozos se pueden usar indistintamente las redes perceptrón multicapa o las máquinas comitè, en función de la complejidad de los mismos. En datos mal clasificados es recomendable utilizar las máquinas comitè.
- c) Es importante en todo momento probar la congruencia de los resultados con los modelos geológicos – petrofísicos conceptuales.

En el caso de datos bien clasificados es importante utilizar la mayor información posible, sin redundar en ella., en el caso que la clasificación se vea desmejorada, es importante utilizar el menor número de variables como sea posible, sin perder de vista la omisión de información valiosa.

Se recomienda dar mayor peso e importancia los registros que respondan a la litofacie que sea estadísticamente más representativa y/o las facies (arcillosas) que sean más fáciles de identificar (arenas limpias o carbonatos)

7.8 Metodología

La Fig. 7.1 resume propuesta metodológica presentada en este trabajo de tesis.

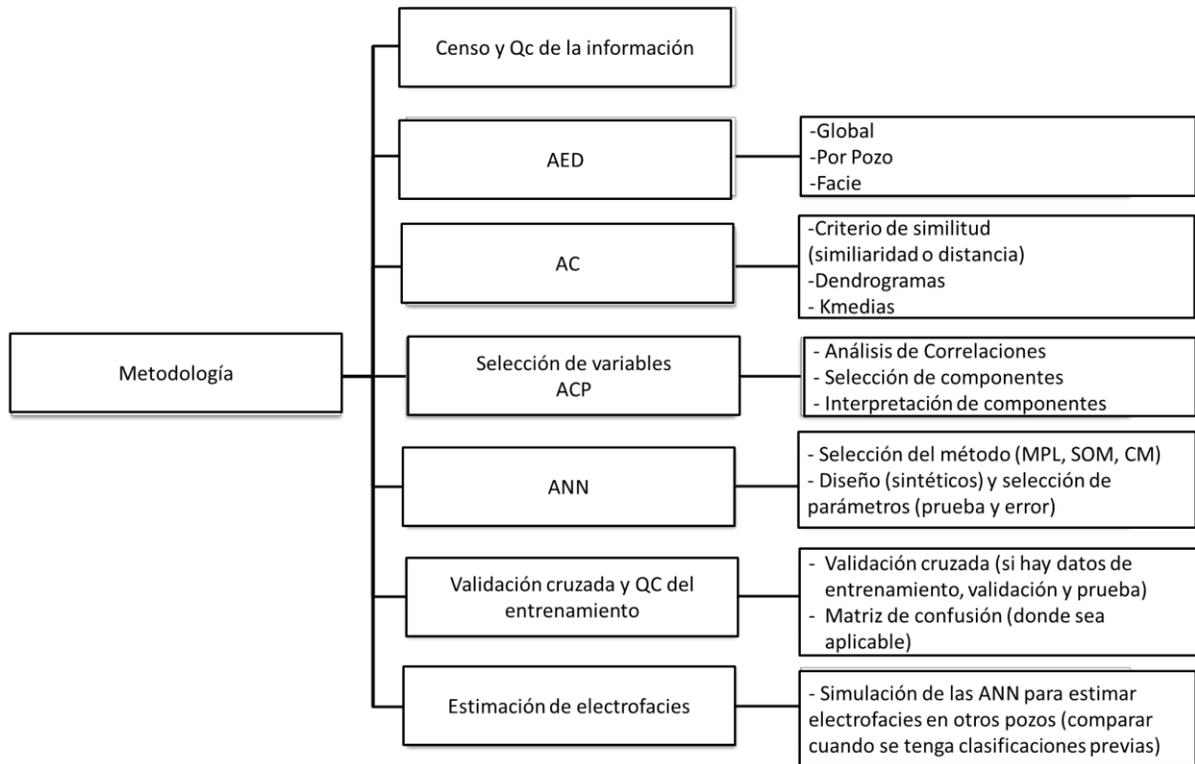


Fig. 7.1 Flujo de la propuesta metodológica para la clasificación de litofacies.

CONCLUSIONES Y TRABAJOS FUTUROS

La metodología plantea uno de los esquemas a seguir dentro de una clasificación automatizada convencional de facies. De los resultados obtenidos en el presente trabajo de tesis, se puede concluir que las redes neuronales son poderosas herramientas de clasificación y su éxito está en función de la naturaleza estadística de los datos.

El flujo de trabajo propuesto tiene como propósito entender el comportamiento de los datos, obtener información para el diseño y entrenamiento óptimo de la red, así como validar los resultados obtenidos de la clasificación mediante uso de la red neuronal. De lo anterior se concluye lo siguiente (de manera desglosada por etapas de la metodología):

1- Criterios de clasificación

Una de las conclusiones que se pudo constatar en este trabajo, es el criterio que debe seguirse, para clasificar un conjunto de datos de registros de pozos, ya que se puede intentar clasificarlos en función de sus patrones de depósito (sistemas siliciclásticos), en función de sus valores de registros (electrofacies) o propiedades petrofísicas y finalmente por su composición litológica. Lo anterior es importante, ya que un determinado tipo de roca, puede tener propiedades petrofísicas diferentes (compactación, porosidad, permeabilidad y saturación de fluidos), lo cual se manifestara en los registros geofísicos de pozo. Como ejemplo supóngase el caso de una arena de canal o levee saturada con agua y un alto grado de compactación, la cual puede tener valores de resistividad similares a un lóbulo no consolidada con impregnación de hidrocarburos, sin en cambio tendrán valores de porosidad neutrón diferentes. En este ejemplo se puede ver que a pesar de ser la misma litología, son facies de depósito diferentes y tiene propiedades petrofísicas diferentes, no obstante tienen valores iguales para la resistividad y diferentes para el registro neutrón, los sistemas automatizados clasificarán en función de los datos de entrada que se propongan, por lo que el conocimiento especializado en la materia (petrofísica y geológica) y en el área de estudio, es de vital importancia para entender los resultados obtenidos. En este trabajo el criterio de clasificación fue el contenido de arcilla, para tratar de definir electrofacies asociadas al litotipo, que serán las guías de la propagación de propiedades petrofísicas, realizada por técnicas como las geoestadísticas, durante el modelado geológico petrofísico de un área.

2- Registros propuestos.

El uso de los registros propuestos RT/GR (utilizado en flujos de trabajo convencionales para identificar secuencias arcillosa y arenosas) y dGR/dZ , tienen como finalidad establecer un parámetro geométrico, que permitiera definir los cambios o inflexiones de la curva de rayos gama, la cual es principalmente la que permite identificar patrones de depósito en función de su forma. En este trabajo fueron descartadas (RT/GR respondió mejor en los registros sintéticos), ya que del análisis exploratorio de datos, se observó, que no proporcionaban información útil, para discriminar las electrofacies.

3- Registros sintéticos.

El uso de registros sintéticos, permitió establecer los parámetros claves de un diseño de red neuronal, como lo son el número de capas, el número de neuronas, el número de elementos en una red de conjunto, así como algunos de los parámetros de los algoritmos de entrenamiento. También permitió ilustrar el comportamiento de una red neuronal ante diferentes conjuntos de datos con características estadísticas diferentes (datos sintéticos de bien a mal clasificados).

4- Análisis Exploratorio de datos.

El análisis exploratorio de datos es una herramienta fundamental para lograr los objetivos de la metodología, ya que con ella se puede prever las dificultades que tendrá un sistema clasificador ante un conjunto de datos determinado. En el caso de los registros sintéticos, el análisis exploratorio de datos (AED), dejó prever la problemática de clasificar el modelo tres, entrenando la red neuronal supervisada (MPL) con los registros sintéticos de los modelos uno y dos; lo anterior se debe a que una red neuronal responde a datos que ya conoce (datos con los que fue entrada), por lo que es necesario entrenarla con información consistente, equivalente y que responda a las electrofacies que se están proponiendo, sea cual sea su naturaleza.(electrofacies petrofísicas, litofacies, facies de depósito, etc.). El análisis exploratorio de datos constató la calidad de la clasificación propuesta por el grupo de especialistas del IMP, en los pozos del campo en estudio, permitió entender la consistencia de las electrofacies en función de los datos y la consistencia de la clasificación entre todo el conjunto de pozos.

5- Análisis de Grupos

El análisis de grupos, es una herramienta alterna que explota las propiedades de correlación entre las diferentes variables, para clasificarlas. En el caso de los registros sintéticos, el análisis de k medias respondió a la clasificación real de los datos. En el caso de los pozos reales, la clasificación con los modelos neuronales superó por mucho a la técnica de las k-medias, aproximándose en un grado aceptable a la clasificación propuesta por el grupo de expertos. El uso de dendrogramas permitió prever el número de grupos que se podría formar, información útil al momento de interpretar los resultados de las redes no supervisadas (SOM), tratando de buscar el menor número de grupos, para facilitar la interpretación de los datos.

6- Análisis de componentes principales.

El análisis de componentes principales, es una técnica que permitió generar criterios de decisión, al momento de escoger las variables o registros de pozos más útiles para entrenar el modelo neuronal.

7- Modelos Neuronales

Los diferentes modelos neuronales propuestos respondieron de forma muy similar. El grado de clasificación con respecto a las electrofacies reales (sintéticos) o propuestas (datos de pozo real), fue mejor para las máquinas comitè (CM), seguido de las redes perceptrón multicapa (MLP) y finalmente los mapas auto-organizados (SOM). La capacidad de subdividir tareas permitió que las unidades individuales que lo conforman realicen la tarea de manera más óptima, además de tener un esquema de entrenamiento muy bien definido, donde los elementos clave son el sobre entrenamiento y la redundancia de los diferentes elementos o sistemas expertos para una subtarea asignada, así como también los criterios de decisión, que permiten explorar un amplia gama de soluciones, para que al final se pueda tomar una solución final (ya sea compuesta o una solución ganadora).

Las redes perceptrón multicapas tienen buen funcionamiento, sin embargo, el trabajo lo realiza únicamente un sistema experto. Los mapas auto-organizados, tienen el inconveniente de confundir el conjunto de datos que se encontraban en los límites que definen cada una de las electrofacies, y como se pudo constatar en el análisis exploratorio de datos, estos límites son difusos, por lo que para trabajos posteriores se propone incorporar el uso de elementos de semántica de lógica difusa.

8- Resultados

De los resultados obtenidos en los pozos del campo en estudio, se puede observar que principalmente existe una consistencia con las cuatro litofacies propuestas. La clasificación que comúnmente manejan los grupos de geociencias de PEMEX Exploración y Producción es principalmente de tipo depositacional, identificando tres principales grupo de facies (ver capítulo 4). En este trabajo, el uso de la clasificación propuesta tiene objetivos más computacionales, enfocado principalmente al modelo estático de un yacimiento.

Los sistemas automatizados que se aplicaron en este estudio, permiten identificar principalmente tres tipos de secuencias (Secuencia como patrón de depósito, no como unidad estratigráfica, la cual requiere datos paleontológicos), donde cada una a su vez define un litotipo para cada punto del registro. Los resultados mostrados, dejan ver los criterios que utilizaron los especialistas para definir su clasificación, la cual está basada principalmente los registros de rayos gama y resistividad; y es por ello que dándole mayor peso a estos registros las redes neuronales reproducen prácticamente la clasificación propuesta por los especialistas. Sin embargo, si se diera más peso a los registros que responden a la porosidad o al contenido de fluidos, se tendrían resultados parecidos a la clasificación alterna que realizaron los expertos del IMP, cuyo criterio es petrofísico.

En resumen las redes neuronales en este trabajo no solo sirvieron de estimadores, sino que permitieron darle validez a la consistencia de la clasificación propuesta por los especialistas; un ejemplo de ello son los resultados del pozo 318, que tiene un comportamiento diferente a los pozos del arreglo de la macropera, el cual fue identificado por los diferentes modelos neuronales, sin embargo, de manera general se pudieron identificar las mismas secuencias de depósito.

De los resultados finales se concluye que la secuencia más arenosa, está conformada principalmente por rasgos que tienden más a ser lóbulos de arena, en lugar de facies canalizadas, lo cual deberá corroborarse mediante el amarre sísmico y el estudio de atributos sísmicos en futuros trabajos.

9- La importancia de las redes neuronales en la exploración de hidrocarburos y la importancia de un modelo integral de facies.

El flujo de trabajo presentado en esta tesis, tiene como punto medular el conocimiento y entendimiento de los datos, para desarrollar una metodología de fácil aplicación y así explotar las bondades de los sistemas neuronales (tecnología catalogada dentro de los sistemas de inteligencia artificial), las cuales son ampliamente utilizados en diferentes campos del conocimiento y aplicadas en diferentes áreas de las geociencias en la industria petrolera.

Este trabajo intenta generar conciencia sobre el uso de tecnologías emergentes, ya que son herramientas que han venido a facilitar el trabajo del explorador y han revolucionado los métodos convencionales de estimación, simulación y clasificación de datos en las geociencias, sin embargo, como cualquier sistema computacional, la red neuronal no ésta exenta de errores y a malas clasificaciones. El uso de datos no adecuados arrojará resultados no adecuados. En este contexto conocer el dato antes del entrenamiento es vital, para obtener un resultado satisfactorio, ya que muchas veces son entrenadas con datos que tiene un alto grado de dependencia lineal entre ellos, ya que proceden de un mismo dato de origen, por citar un ejemplo clasificación de forma de onda en estudios sísmicos multiatributos (aplicaciones como Stratimagic, Petrel, Opendetec, etc., tienen módulos especializados).

Un modelo correcto de facies, ya sean de depósito, litofacies o electrofacies petrofísicas, es muy importante para entender entre otras cosas los modelos geológicos integrales (modelo de depósito) de

un área prospectiva, con el fin de comprobar la existencia de los elementos del sistema petrolero y la calidad de la roca almacén, para aumentar la probabilidad de éxito geológico y poder documentar posibles prospectos a perforar. Un modelo de facies de vital importancia también para la propagación de propiedades petrofísicas en un modelo geológico petrofísico, durante la caracterización estática, el cual es el insumo principal para la caracterización dinámica de un yacimiento, la certificación de reservas y los planes de desarrollo de un campo potencial. Es por ello que la integración de toda la información es sumamente importante, por ello se proponen para futuros trabajos, el uso de las redes neuronales con información sísmica y pozos, utilizando el mismo criterio planteado en esta tesis.

La información sísmica, en especial los atributos sísmicos, son herramientas importantes, que junto con el mapeo de la forma de onda sísmica y la interpretación, permiten inferir los sistemas de aporte en una cuenca, los cuales quedan definidos por las facies sísmicas. Las facies sísmicas son en realidad quienes guiarán el comportamiento petrofísico en un modelo de yacimiento, por tal motivo es necesario conocer que tipos de atributos sísmicos se utilizarán, tratando de evitar la alta correlación lineal entre ellos, (la fase, la amplitud y la frecuencia son propiedades sísmicas independientes entre sí), buscando de preferencia aquellos cuya respuesta está ligada al comportamiento petrofísico de un tipo de roca. Por lo que se propone el análisis estadístico multivariado de atributos como: la descomposición espectral, atributos de amplitud, de fase, de frecuencia y coherencia, para la clasificación de la forma de onda mediante redes neuronales. Si se cuenta con estudios de inversión sísmica, se propone el uso de la impedancia acústica (IA), como la conexión entre las facies estimadas en el pozo y las sismofacies (algunos software como Geoview de Hampson and Rusell, siguen este esquema). Todo lo anterior siguiendo la metodología basada en el análisis exploratorio de datos para clasificar mediante redes neuronales.

LISTA DE ABREVIATURAS

- AA*: Análisis de agrupamiento o de clústeres
- ACP*: Análisis de componentes principales.
- AED*: Análisis exploratorio de datos (DEA- Data Exploration Analysis)
- AIC*: Criterio de Información de Akaike (Akaike Information Criteria)
- ART*: Teoría de la Resonancia Adaptativa (Adaptive Resonance Theory)
- BAM*: Memoria Asociativa Bidireccional (Bidirectional Associative Memory)
- BP*: Backpropagation (Algoritmo de retropropagación)
- BPANN*: Red Neuronal Backpropagation (Backpropagation Artificial Neural Network)
- CAL*: Acrónimo del registro caliper
- CM*: Máquina comité (Comité Machines)
- dGR/Dz: Derivada del rayos gama con respecto a la profundidad.
- DT*: Acrónimo, del registró sónico- tiempo de tránsito
- GR*: Acrónimo del registro de rayos gama
- LMBP*: Algoritmo Levenderg-Marquardt BP
- LMS*: Mínimos cuadrados (least Mean square)
- MGPY*: Modelación Geológica Petrofísica de Yacimientos
- MLP*: Red Perceptrón Multicapa (Multi Layer Perceptrón)
- MSE*: Error cuadrático medio (Mean Square Error)
- NPHI*: Acrónimo del registro de porosidad neutrón
- OLC*: Combinación lineal óptima (Optimal Lineal Combination)
- RHOB*: Acrónimo del registro de densidad
- RNA*: Red Neuronal Artificial (ANN en inglés Artificial Neural Network)
- RT*: Acrónimo del registro de resistividad de la formación
- RT/GR*: Relación Resistividad – Rayos Gama
- SOM*: Mapas auto-organizados (Self Organize Map)
- TA*: Tasa de Aprendizaje
- VC*: Dimensión de Vapnik-Chervonenkis

Apéndice A

Análisis Exploratorio de Datos (AED) Registros Sintéticos

I. AED por clase

- Estadísticos por clase.

Tabla A.1 Estadísticos clase uno

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	59.37	2.01	7.283	1.434	70.429	0.275	0.0254	-35.2457
1er Cuartil	67.4585	2.1793	7.8243	2.035	85.8962	0.3048	0.0301	-3.7515
Mediana	74.614	2.8825	8.044	2.1655	102.2465	0.352	0.0402	0.0328
Media	73.7138	3.1802	8.006	2.1282	98.5128	0.343	0.0411	-0.9436
3er Cuartil	79.883	4.0672	8.181	2.2363	112.0408	0.3785	0.0518	2.2842
Máximo	84.949	5.951	8.665	2.643	122.189	0.411	0.06	15.8453
Rango	25.579	3.941	1.382	1.209	51.76	0.136	0.0346	51.091
Rangos Intercuartil	12.4245	1.888	0.3567	0.2012	26.1445	0.0737	0.0218	6.0357
Varianza	48.9501	1.1799	0.0763	0.0275	219.494	0.0016	0.0001	57.3844
Desviación Estándar	6.9964	1.0862	0.2762	0.1659	14.8153	0.0405	0.0112	7.5752
Simetría	-0.3676	0.4784	-0.2688	-0.7711	-0.1629	-0.033	0.0869	-1.6378

Tabla A.2 Estadísticos clase dos

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	79.501	1.369	7.244	1.64	91.001	0.33	0.0123	-32.44
1er Cuartil	97.9908	1.4863	8.5503	1.8752	101.6325	0.364	0.0143	-5.3375
Mediana	102.259	1.5505	8.8615	1.939	138.7755	0.43	0.0152	0.1781
Media	101.7527	1.6732	8.8307	1.9457	127.5383	0.4154	0.0166	0.662
3er Cuartil	105.6167	1.851	9.1875	2.0107	145.529	0.4548	0.0185	4.9821
Máximo	122.157	2.42	10.027	2.362	157.212	0.494	0.0271	42.44
Rango	42.656	1.051	2.783	0.722	66.211	0.164	0.0148	74.88
Rangos Intercuartil	7.626	0.3647	0.6372	0.1355	43.8965	0.0908	0.0042	10.3196
Varianza	52.0679	0.0683	0.2876	0.0127	492.7592	0.0023	0	119.4575
Desviación Estándar	7.2158	0.2612	0.5363	0.1129	22.1982	0.0477	0.0032	10.9297
Simetría	-0.4024	0.9983	-0.4138	0.2164	-0.4162	-0.3007	1.181	0.4612

Tabla A.3 Estadísticos clase tres

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	60.234	1.351	7.179	1.218	85.102	0.252	0.0139	-39.12
1er Cuartil	76.8988	1.9968	7.9577	1.8508	93.29	0.335	0.025	-7.5975
Mediana	80.0395	2.3675	8.3065	2.0915	111.6465	0.35	0.0276	-0.3474
Media	85.0292	2.7794	8.4235	2.0133	115.9341	0.3649	0.0323	-0.7722
3er Cuartil	93.8718	3.3727	8.858	2.1873	139.0175	0.3885	0.0431	6.1295
Máximo	120.628	5.886	10.142	2.46	152.425	0.484	0.0607	45.31
Rango	60.394	4.535	2.963	1.242	67.323	0.232	0.0468	84.43
Rangos Intercuartil	16.973	1.376	0.9003	0.3365	45.7275	0.0535	0.0181	13.727
Varianza	165.6514	1.3765	0.4697	0.0543	532.8694	0.0023	0.0001	208.3328
Desviación Estándar	12.8706	1.1733	0.6853	0.2331	23.084	0.0481	0.0122	14.4337
Simetría	0.7458	1.084	0.496	-0.6972	0.1145	0.7263	0.3953	0.1825

- Histogramas por clase

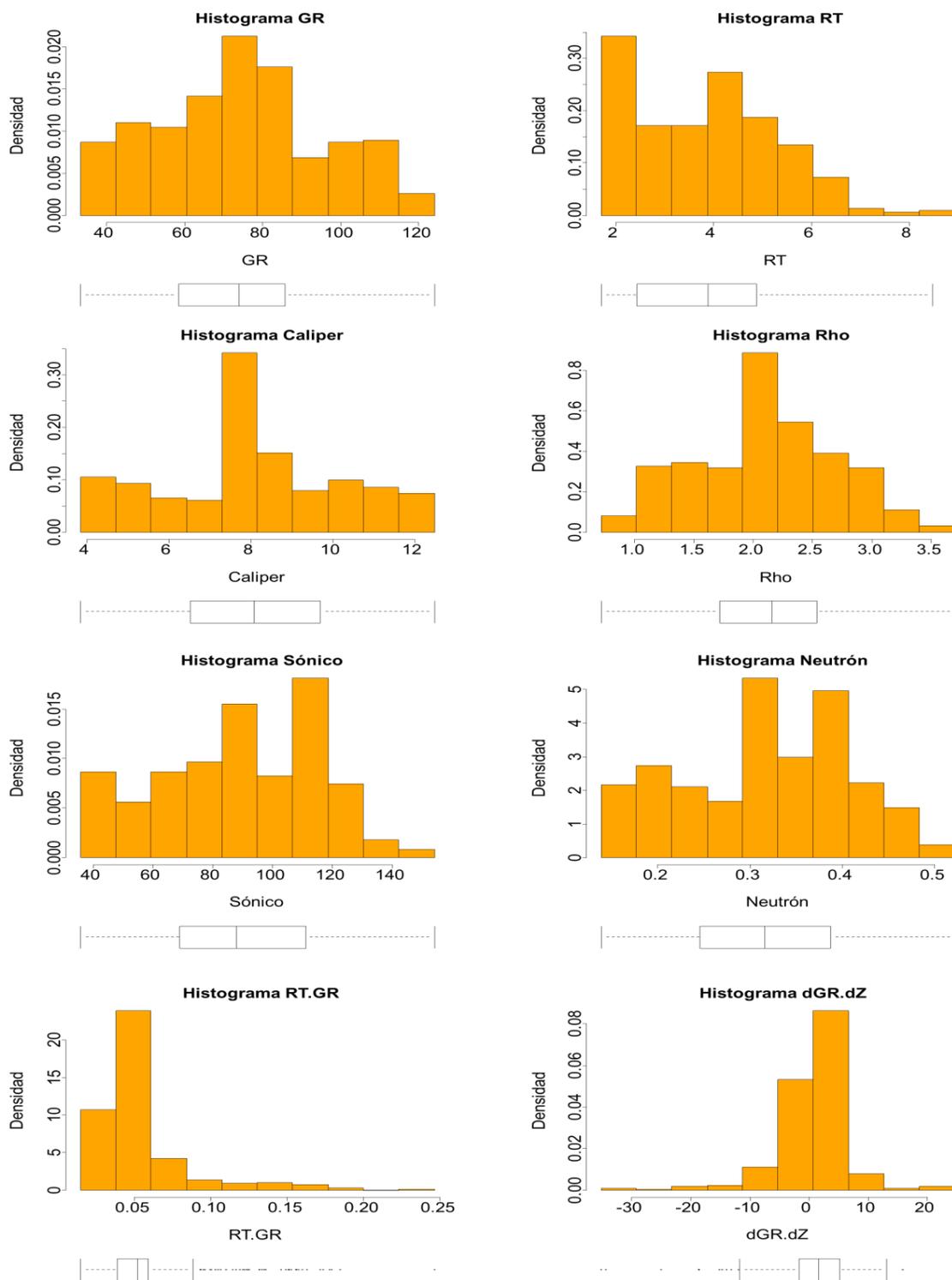


Figura A.1 Histogramas clase uno

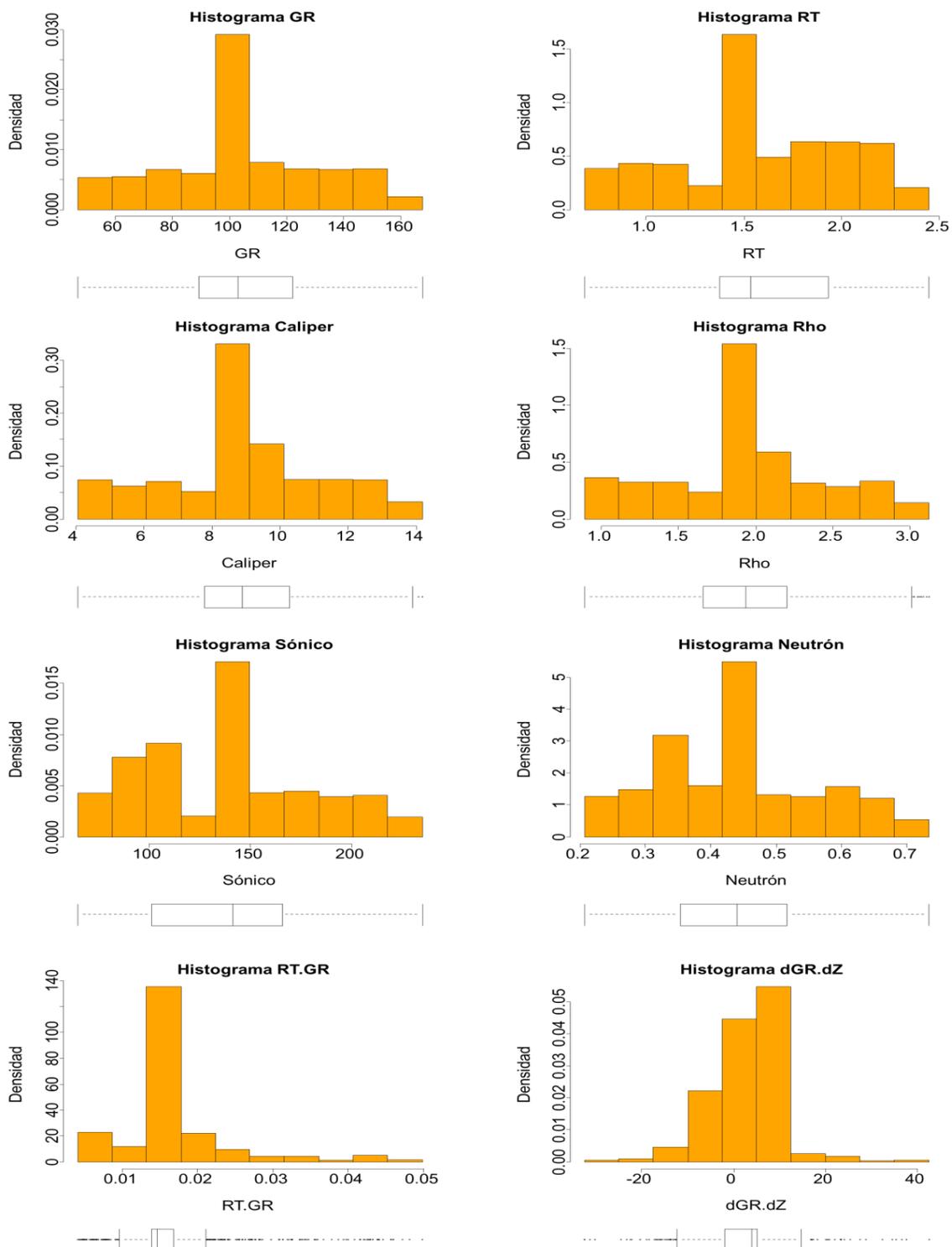


Figura A.2 Histogramas clase dos

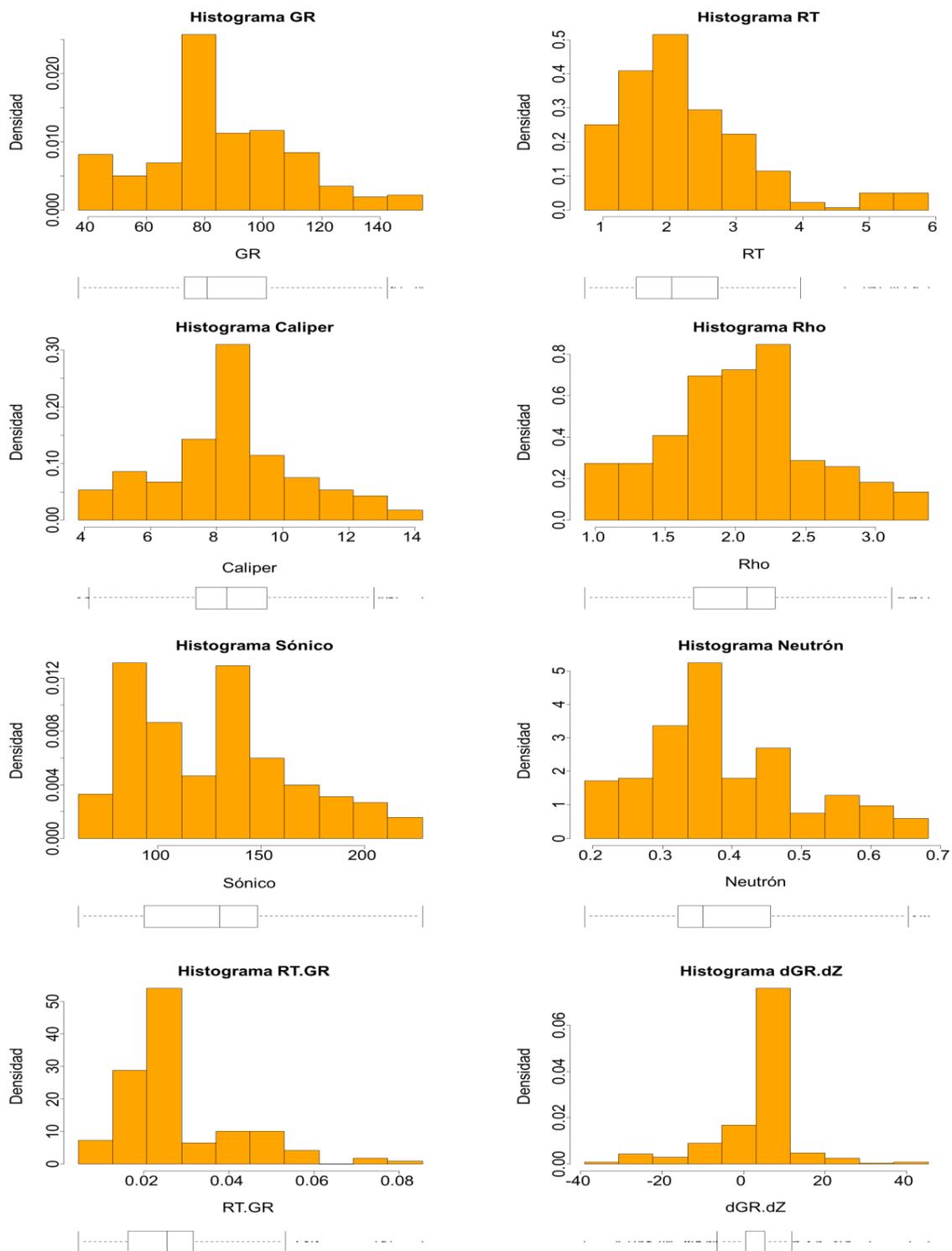


Figura A.3 Histogramas clase 3

- Gráficos cuantil - cuantil por clase

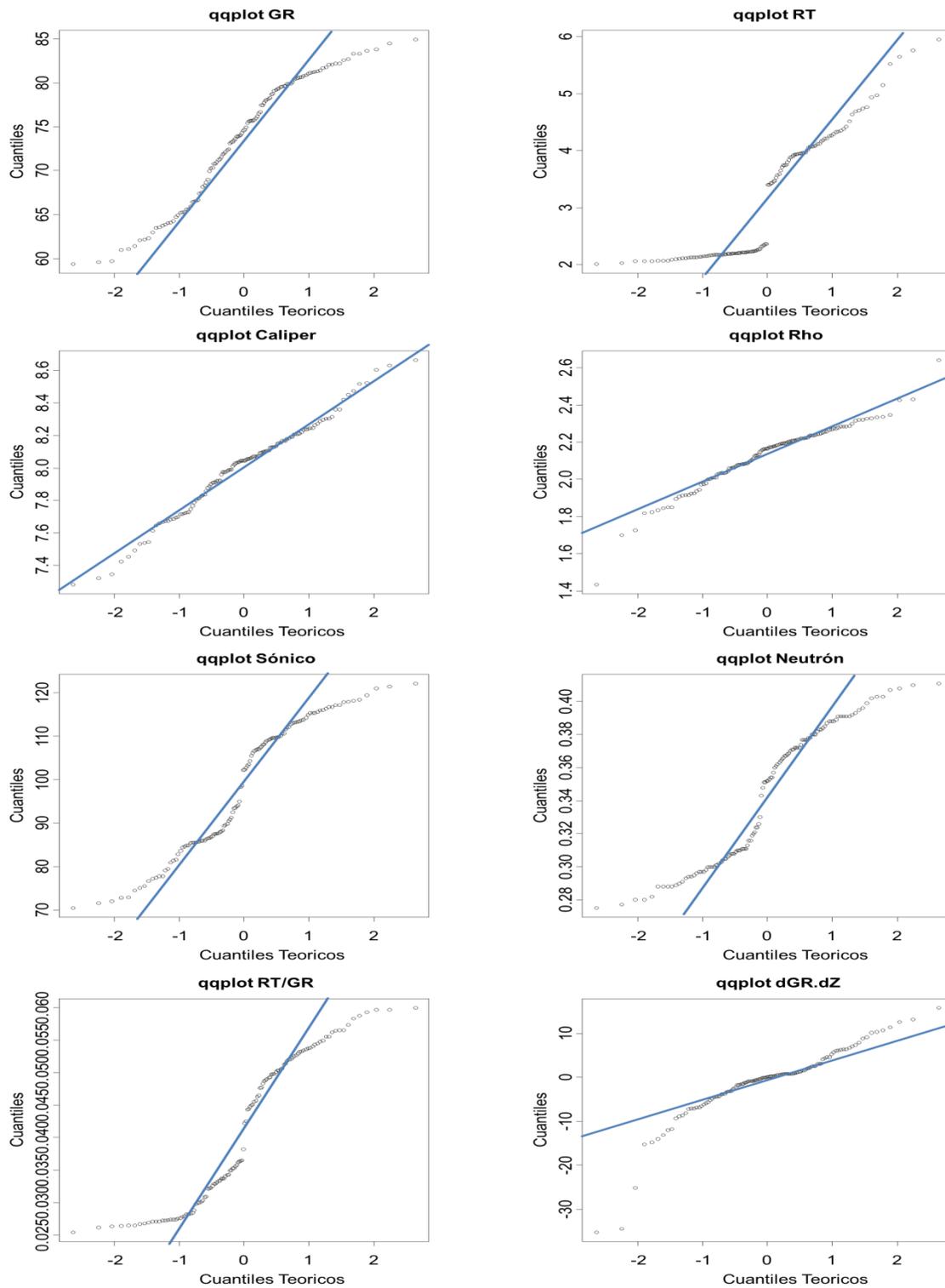


Figura A.4 Gráficos cuantil- cuantil clase uno

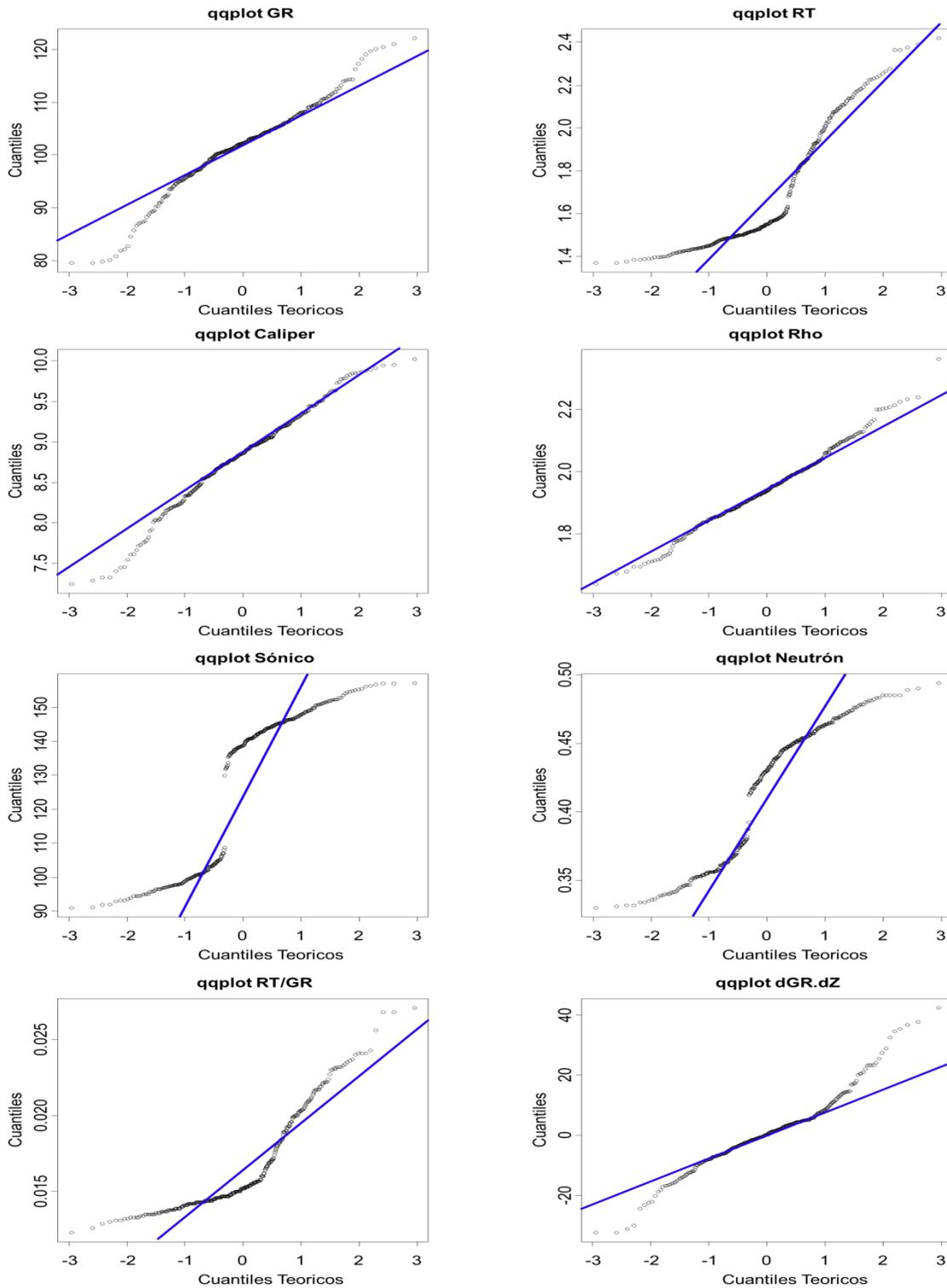


Figura A.5 Gráficos cuantil- cuantil clase dos

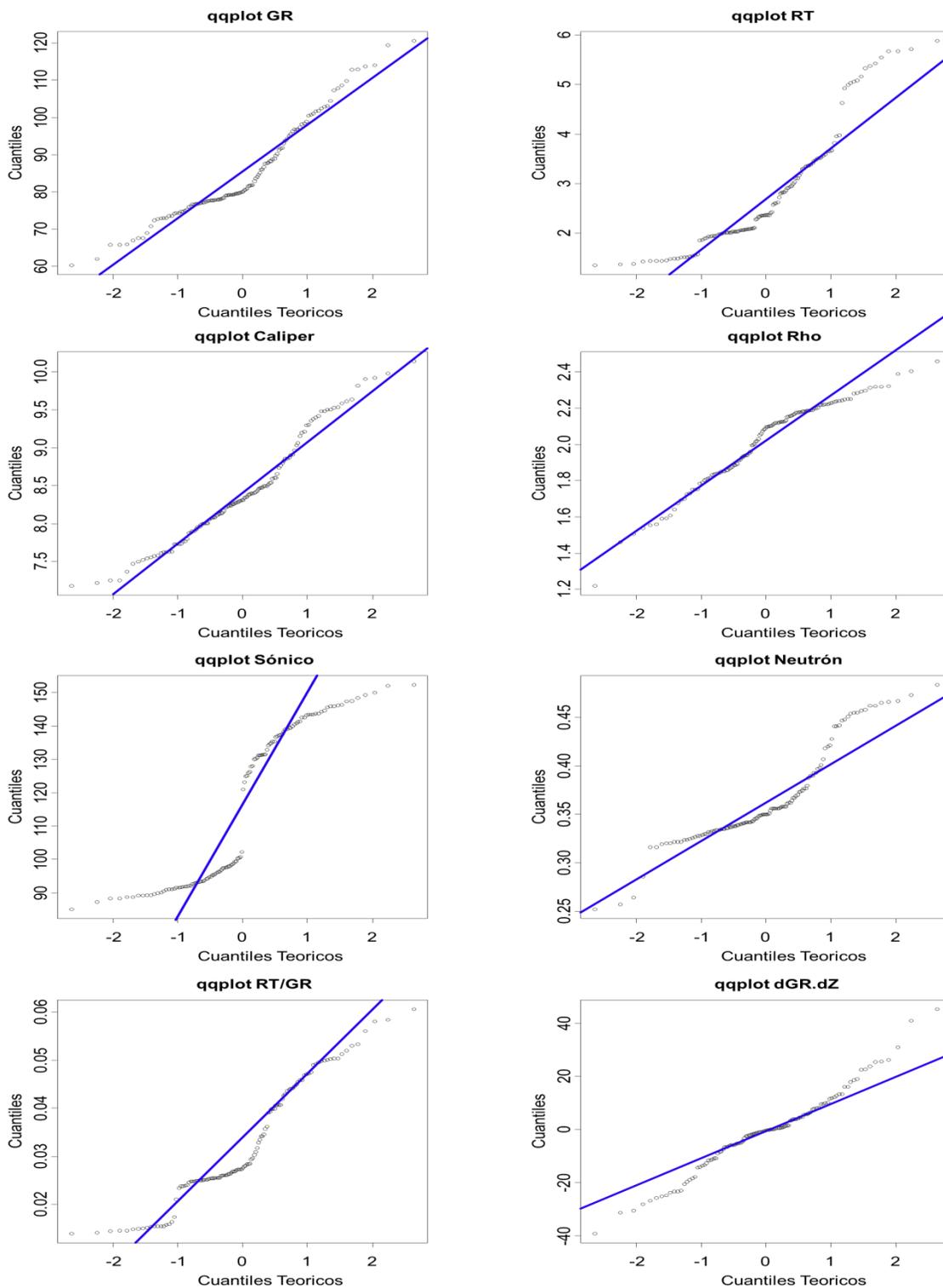


Figura A.6 Gráficos cuantil- cuantil clase 3

- Gráfico de caja por clase

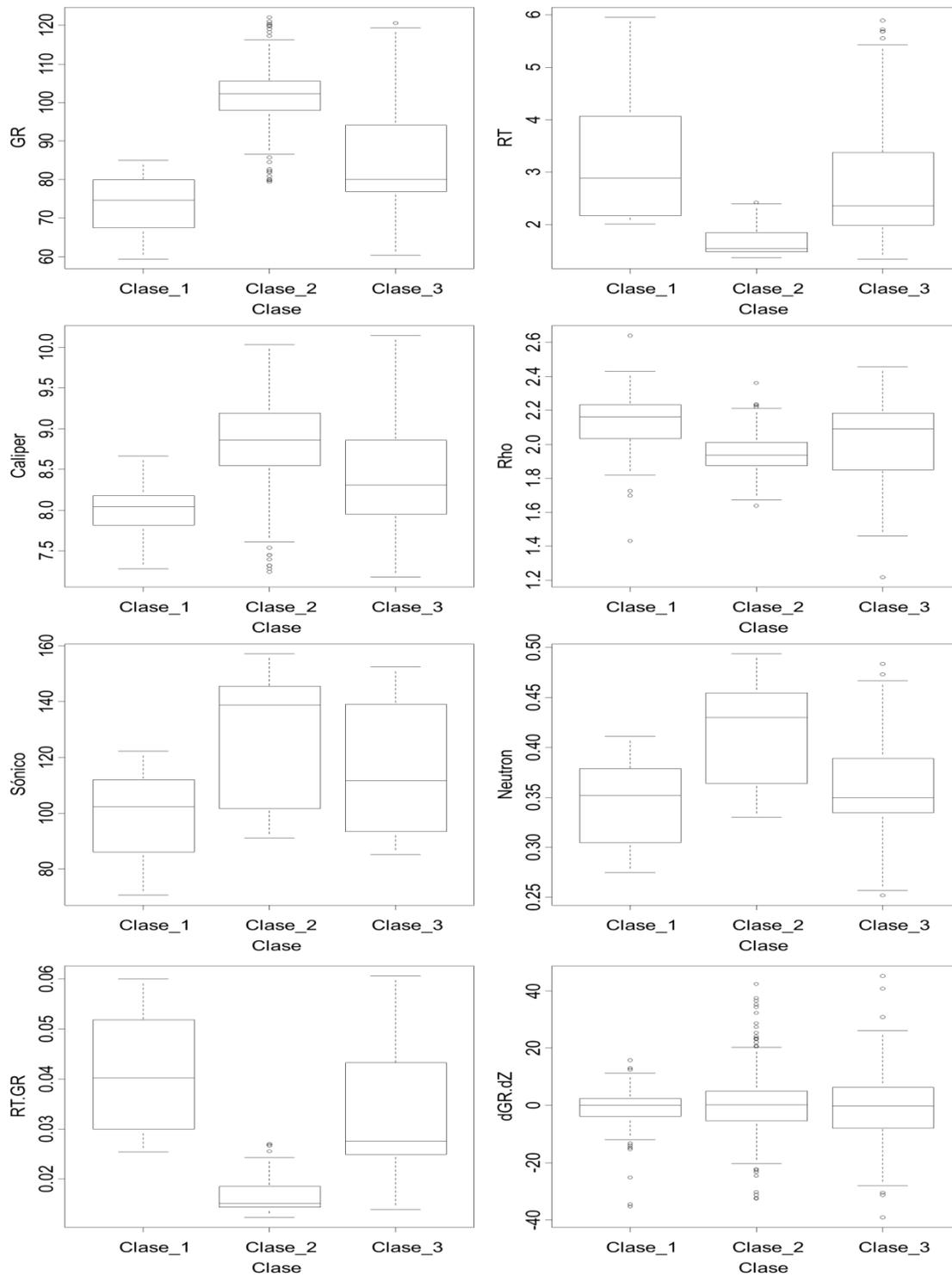


Figura A.7 Gráficos de caja por clase

- Gráfico de dispersión

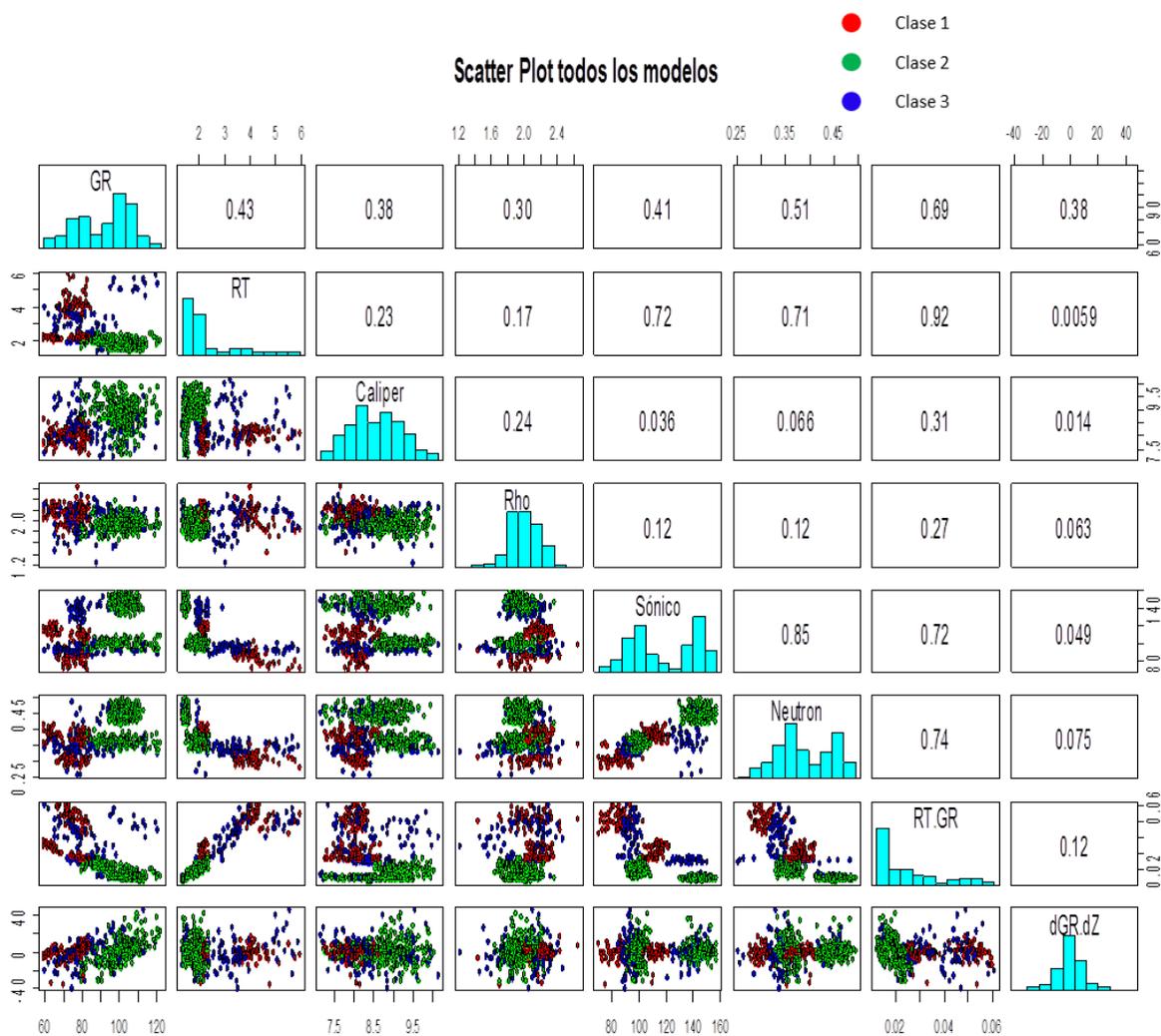


Figura A.8 Gráficos de dispersión de los modelos uno, dos y tres

II. AED por modelo y por modelo-clase

- Gráficos de caja por modelo

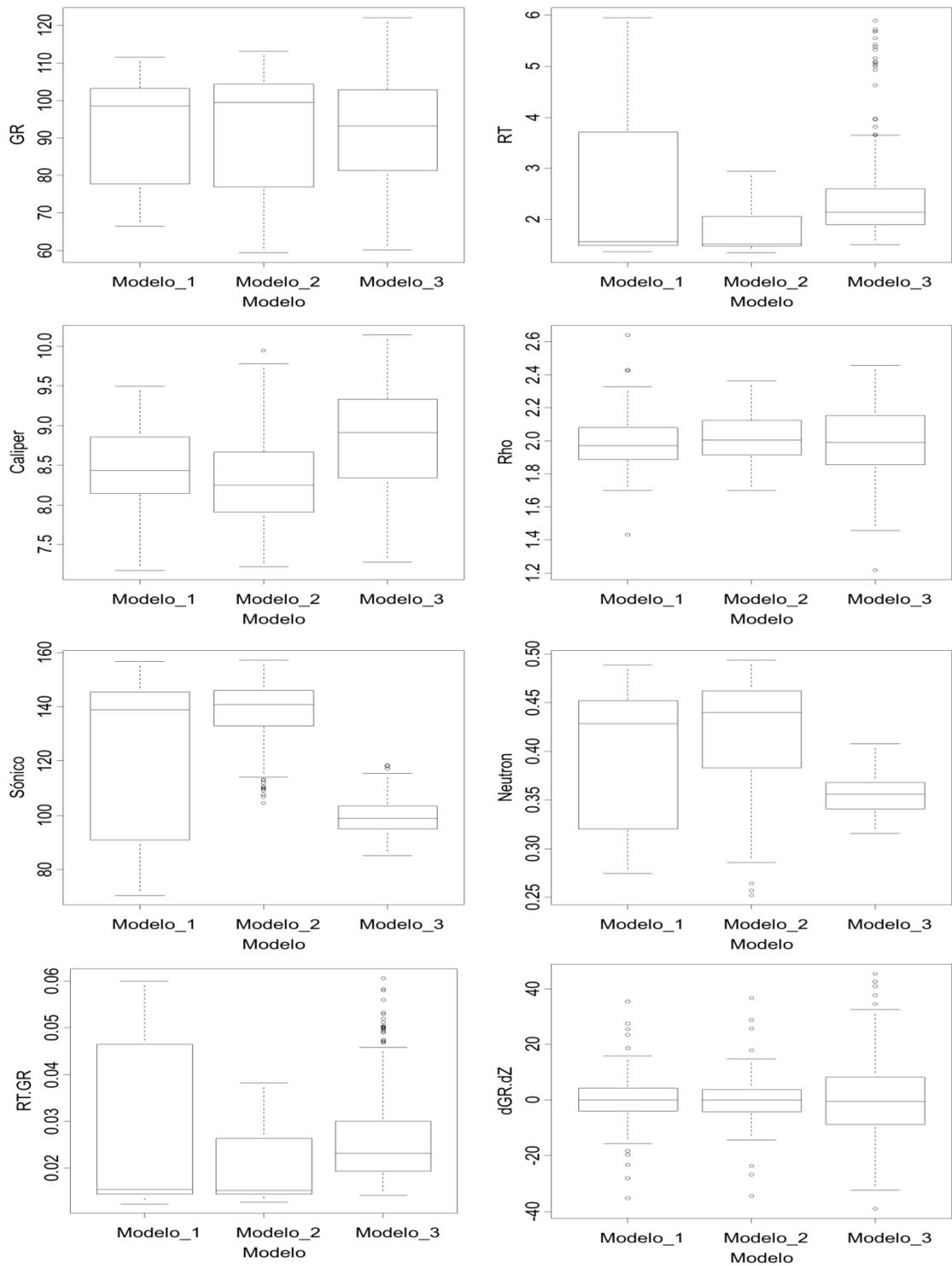


Figura A.9. Gráfico de caja por modelo

A. modelo 1

- Estadísticos

Tabla A.4. Estadísticos generales - modelo 1

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	66.476	1.369	7.179	1.434	70.429	0.275	0.0123	-35.2457
1er Cuartil	77.8197	1.492	8.15	1.8872	90.9387	0.3202	0.0144	-3.9815
Mediana	98.4215	1.5685	8.431	1.971	138.73	0.4285	0.0155	0.135
Media	91.8012	2.3654	8.4844	1.9944	124.9829	0.3989	0.027	0.0079
3er Cuartil	103.201	3.696	8.851	2.0808	145.493	0.452	0.0465	4.293
Máximo	111.64	5.951	9.495	2.643	156.657	0.489	0.06	35.3336
Rango	45.164	4.582	2.316	1.209	86.228	0.214	0.0477	70.5793
Rangos Intercuartil	25.3812	2.204	0.701	0.1935	54.5542	0.1318	0.032	8.2745
Varianza	186.7638	1.5668	0.2116	0.0233	746.7646	0.0044	0.0003	69.0045
Desviación Estándar	13.6662	1.2517	0.46	0.1527	27.327	0.0661	0.0168	8.3069
Simetría	-0.3478	1.0463	0.0923	0.6571	-0.778	-0.5966	0.8139	0.0324

Tabla A.5. Estadísticos modelo 1 – clase uno

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	66.476	3.398	7.674	1.434	70.429	0.275	0.0422	-35.2457
1er Cuartil	71.9467	3.8243	7.9225	1.9245	80.5355	0.2948	0.0488	-5.4095
Mediana	75.2045	4.0685	8.0835	2.069	85.8565	0.3045	0.0519	0.0538
Media	75.4857	4.1803	8.0636	2.0724	85.0507	0.3063	0.0516	-0.9258
3er Cuartil	79.6972	4.3582	8.1947	2.1997	88.6283	0.3115	0.0546	2.6815
Máximo	84.949	5.951	8.36	2.643	102.723	0.357	0.06	15.8453
Rango	18.473	2.553	0.686	1.209	32.294	0.082	0.0178	51.091
Rangos Intercuartil	7.7505	0.534	0.2722	0.2752	8.0928	0.0167	0.0058	8.091
Varianza	21.5994	0.3393	0.0295	0.0415	49.8448	0.0003	0	72.1273
Desviación Estándar	4.6475	0.5825	0.1718	0.2037	7.0601	0.0182	0.0045	8.4928
Simetría	0.1181	1.1834	-0.3576	-0.1549	-0.0875	0.9254	-0.0712	-0.984

Tabla A.6. Estadísticos modelo 1 – clase dos

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	93.686	1.369	8.103	1.761	129.91	0.412	0.0123	-15.5188
1er Cuartil	100.556	1.4488	8.4737	1.8732	139.9765	0.431	0.0141	-2.8399
Mediana	102.844	1.5	8.7775	1.9295	144.2115	0.448	0.0146	0.1781
Media	102.9268	1.5002	8.7548	1.9334	143.9505	0.4465	0.0146	0.8208
3er Cuartil	105.2715	1.551	9.003	1.99	147.0387	0.4575	0.0151	4.0518
Máximo	111.64	1.633	9.495	2.121	156.657	0.489	0.0169	35.3336
Rango	17.954	0.264	1.392	0.36	26.747	0.077	0.0046	50.8524
Rangos Intercuartil	4.7155	0.1022	0.5293	0.1167	7.0623	0.0265	0.001	6.8917
Varianza	14.9268	0.0039	0.1226	0.0064	30.2448	0.0003	0	51.2407
Desviación Estándar	3.8635	0.0623	0.3502	0.0803	5.4995	0.0182	0.0008	7.1583
Simetría	-0.0548	-0.0952	-0.0696	0.1353	0.0889	0.107	0.3228	1.6255

Tabla A.7. Estadísticos modelo 1 – clase tres

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	73.062	1.431	7.179	1.782	121.021	0.356	0.0139	-28.0999
1er Cuartil	77.732	1.5435	8.1368	1.952	130.3558	0.382	0.016	-5.0138
Mediana	79.933	1.976	8.3105	2.0915	135.1165	0.3995	0.0251	-0.2476
Media	84.3798	1.8498	8.3523	2.0575	136.5638	0.4128	0.0225	-1.0244
3er Cuartil	94.0903	2.0263	8.6452	2.1715	143.6927	0.4502	0.0255	4.8083
Máximo	103.049	2.095	9.48	2.25	152.425	0.484	0.0285	25.3267
Rango	29.987	0.664	2.301	0.468	31.404	0.128	0.0146	53.4266
Rangos Intercuartil	16.3583	0.4828	0.5085	0.2195	13.337	0.0682	0.0095	9.8221
Varianza	100.9871	0.0611	0.2682	0.0197	82.136	0.0014	0	126.9963
Desviación Estándar	10.0492	0.2473	0.5179	0.1404	9.0629	0.0373	0.005	11.2693
Simetría	0.815	-0.7724	-0.1961	-0.4859	0.1258	0.2633	-0.7806	-0.3408

• Histogramas

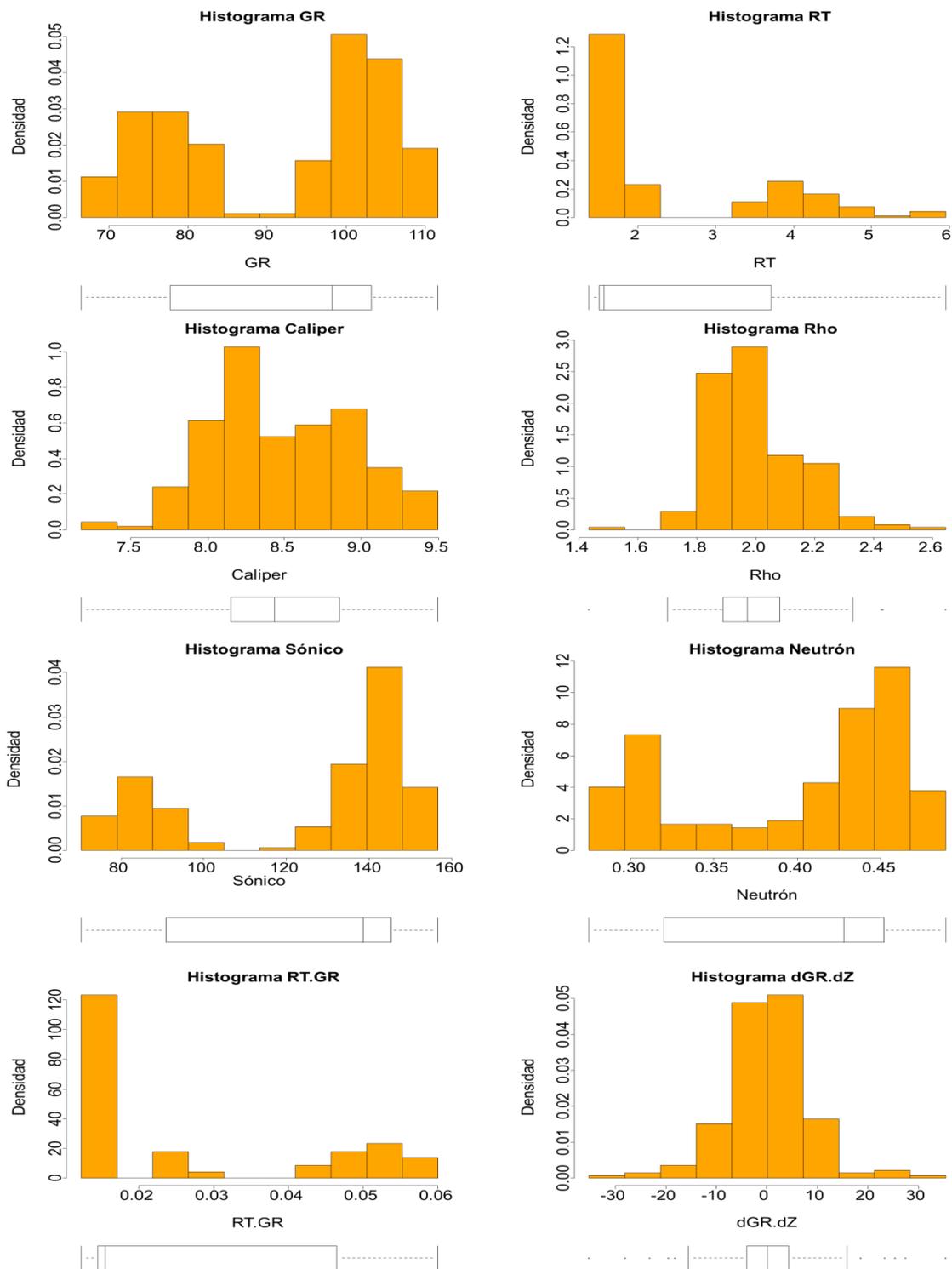


Figura A.10. Histogramas modelo uno

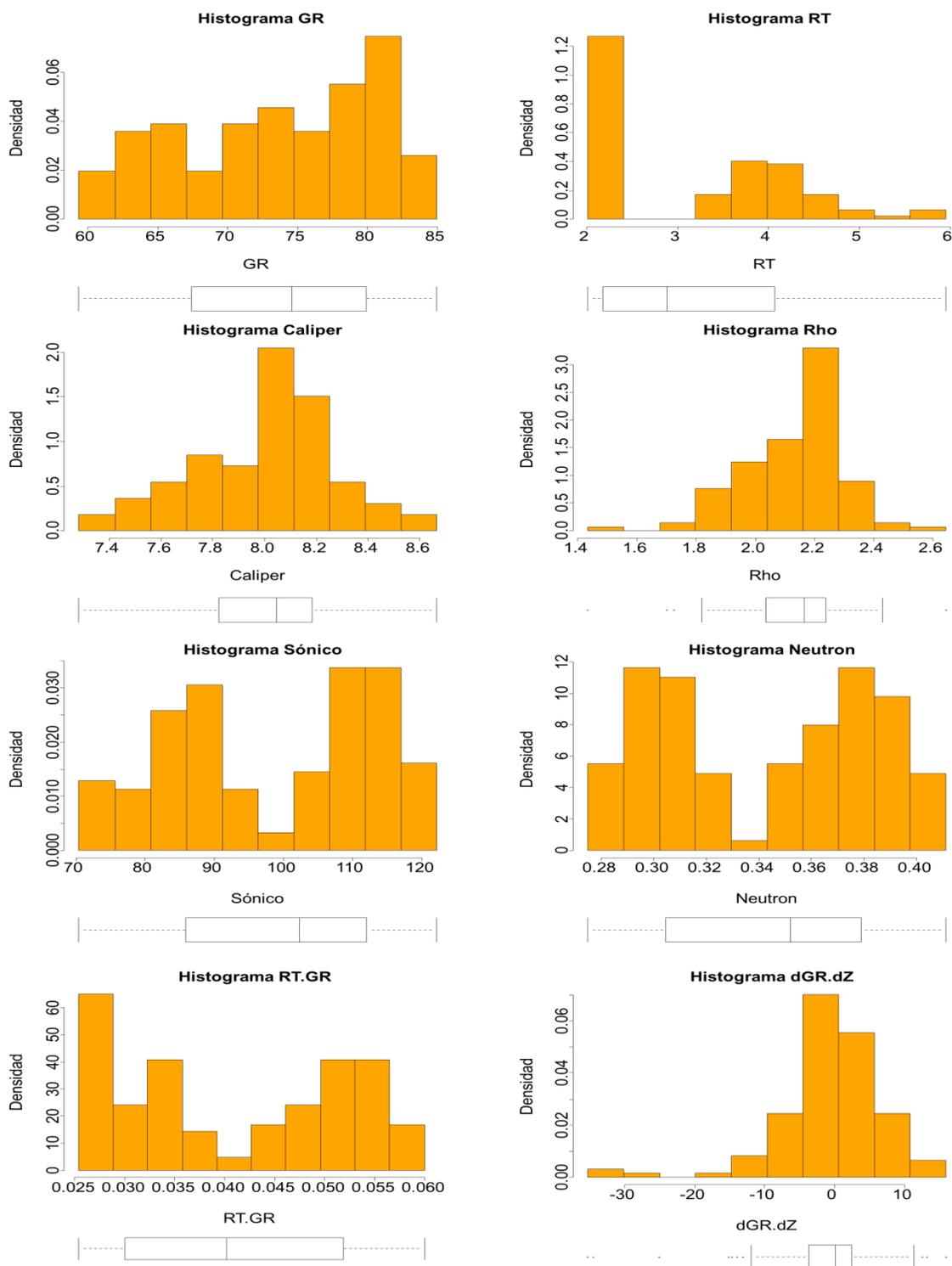


Figura A.11. Histogramas del modelo uno, clase uno

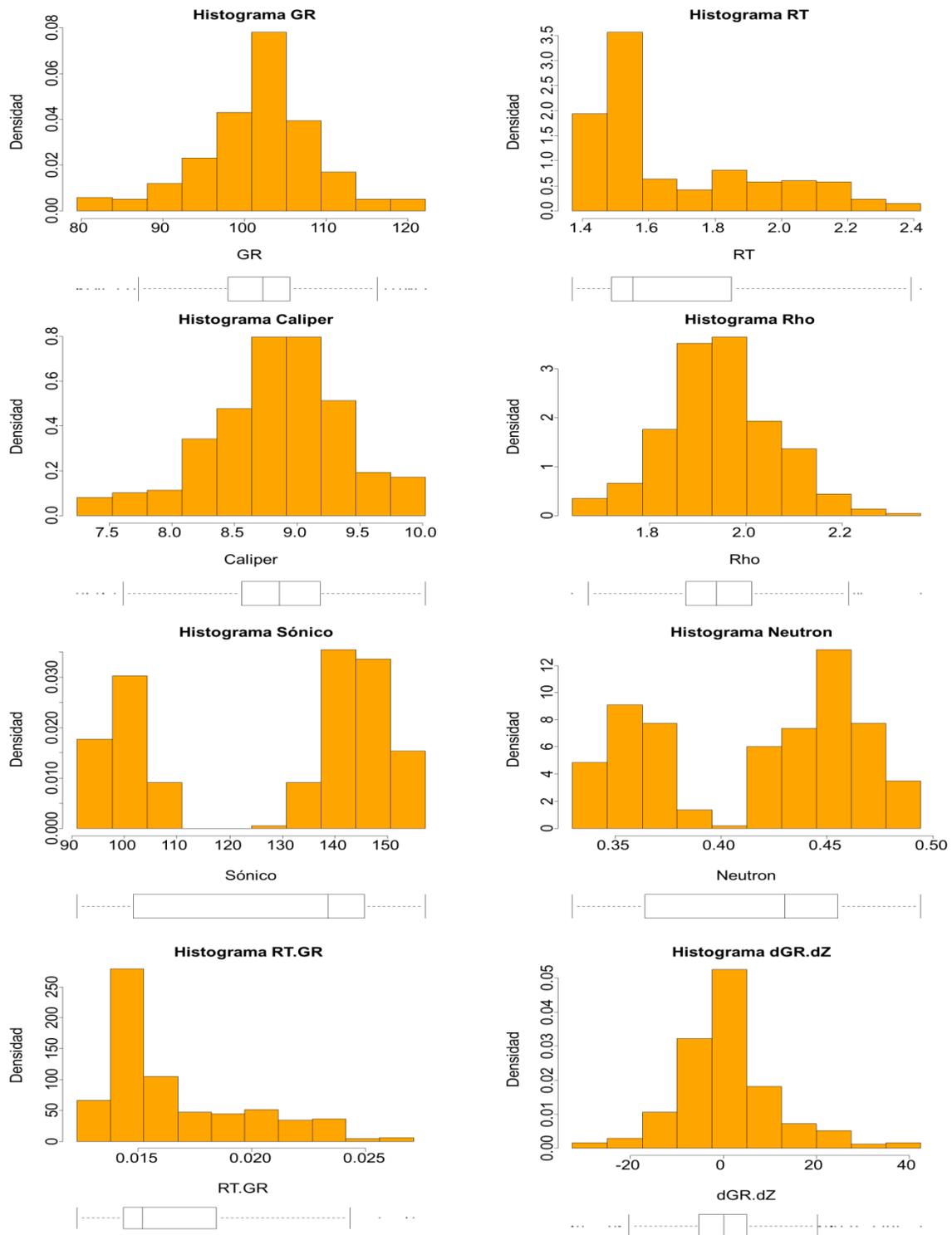


Figura A.12. Histogramas del modelo uno, clase dos

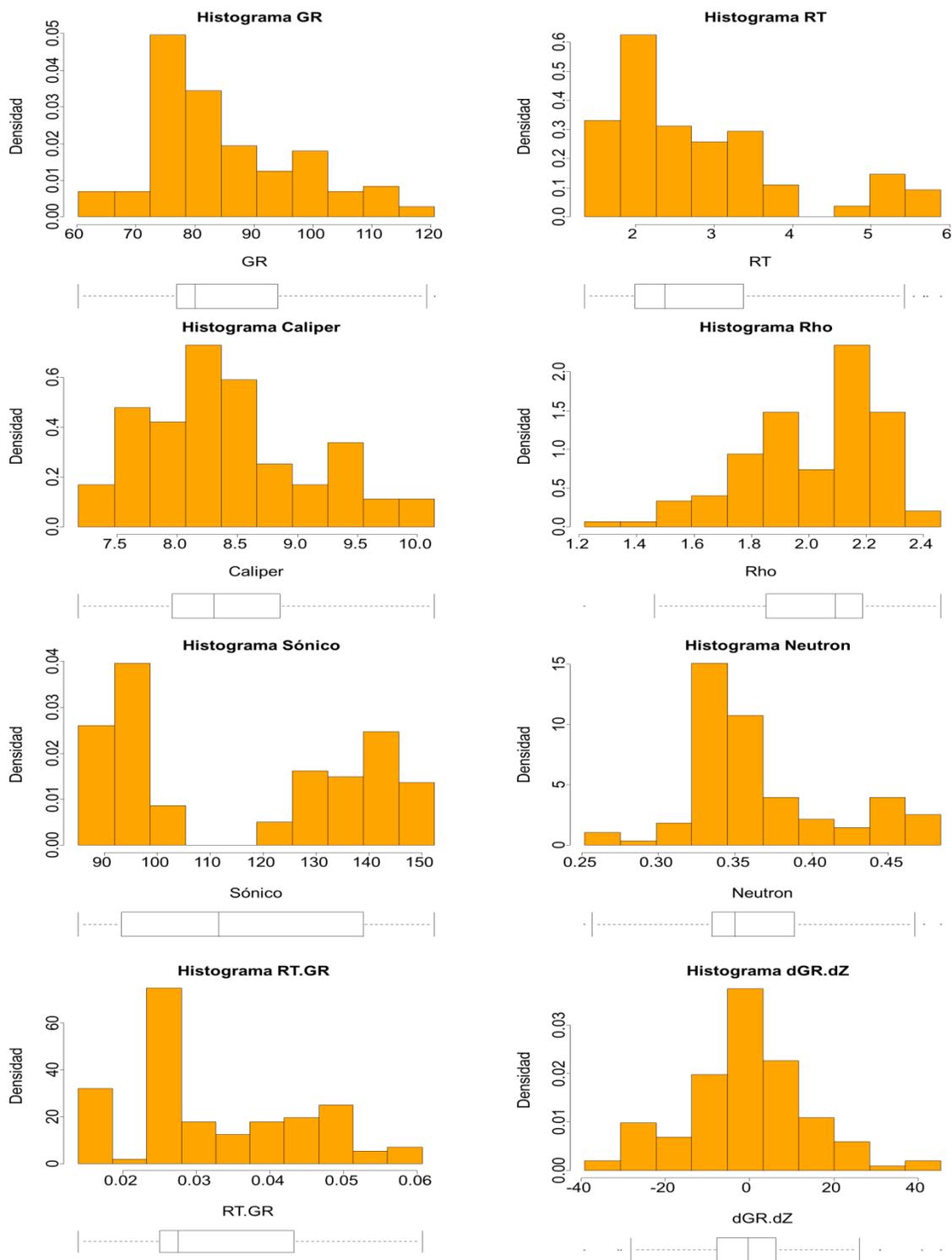


Figura A.13 Histogramas del modelo uno, clase 3

- Gráficos cuantil-cuantil, modelo uno

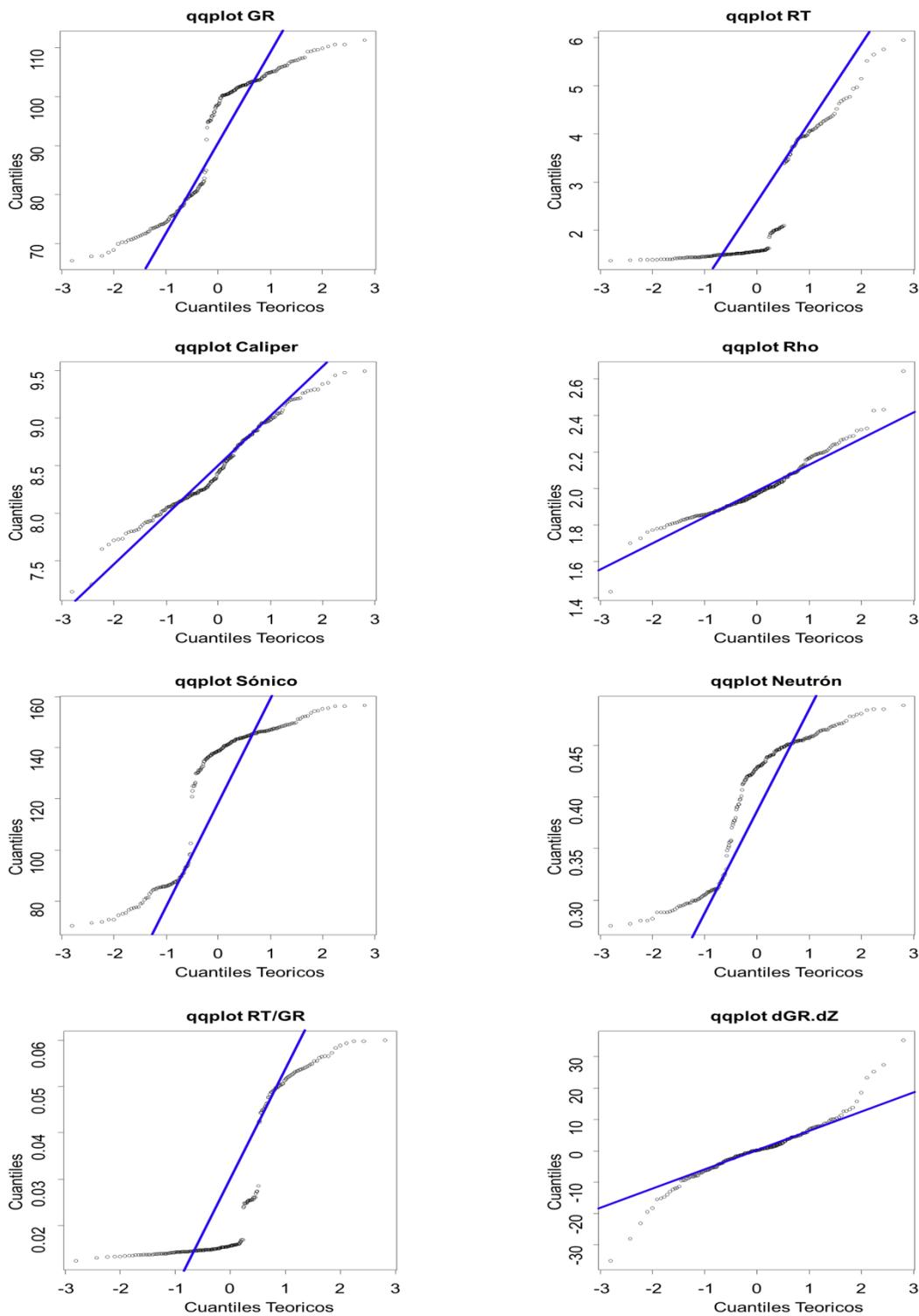


Figura A.14. Gráfico cuantil- cuantil, modelo uno

- Gráfico de dispersión, modelo uno

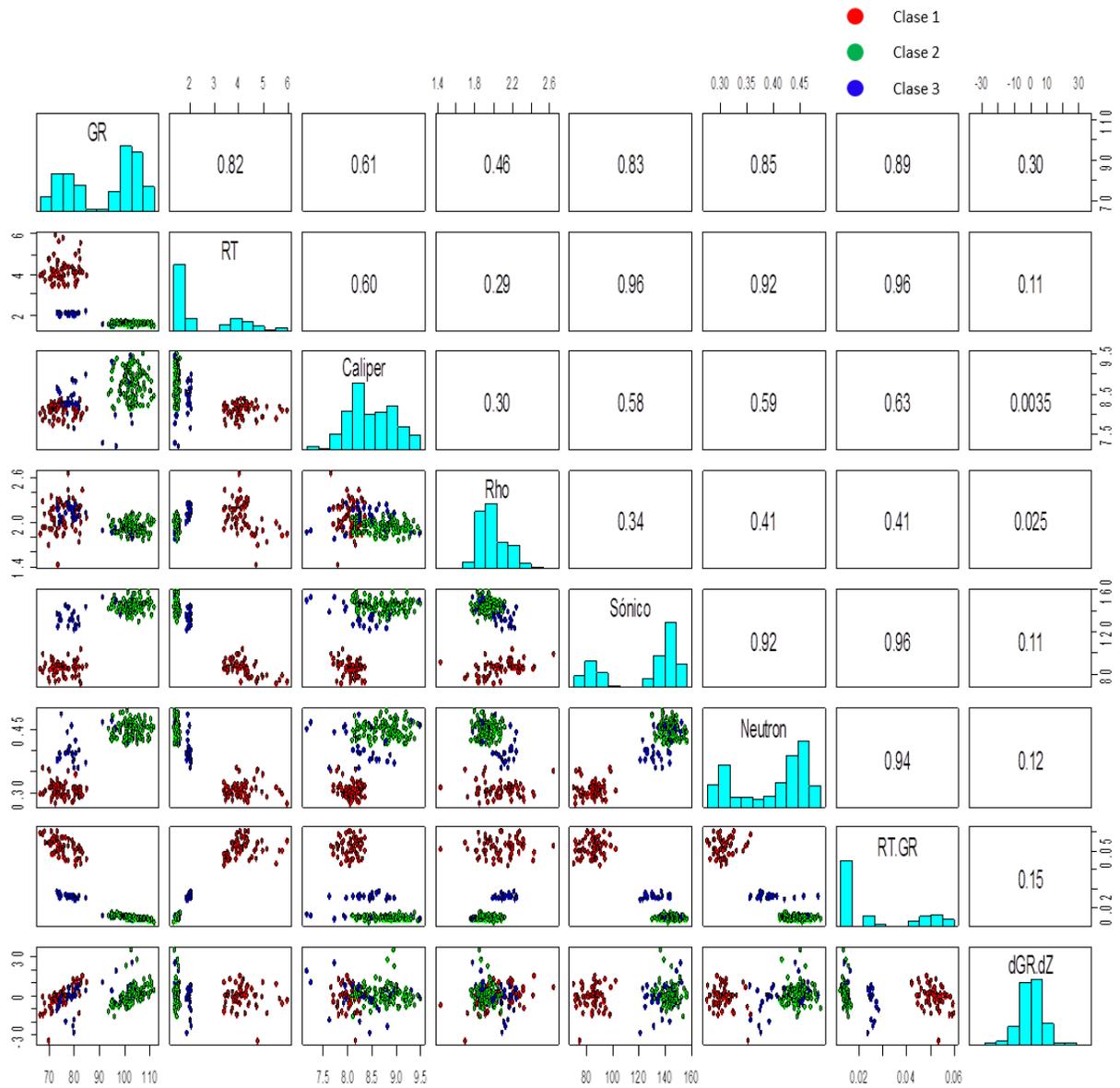


Figura A.15 Gráficos de dispersión modelo 1

B. modelo dos

- Estadísticos

Tabla A.8 Estadísticos modelo dos

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	59.37	1.351	7.217	1.7	104.376	0.252	0.0126	-34.4813
1er Cuartil	76.941	1.48	7.913	1.9155	133.2413	0.383	0.0144	-4.178
Mediana	99.5145	1.527	8.2535	2.0055	140.974	0.44	0.0152	0.093
Media	91.2941	1.7228	8.3125	2.0223	137.6334	0.4222	0.02	-0.0148
3er Cuartiles	104.2818	2.0675	8.6617	2.1227	146.0032	0.462	0.0262	3.6711
Máximo	113.168	2.941	9.944	2.362	157.212	0.494	0.0382	36.6439
Rango	53.798	1.59	2.727	0.662	52.836	0.242	0.0256	71.1252
Rangos Inter cuartil	27.3407	0.5875	0.7487	0.2072	12.762	0.079	0.0118	7.8491
Varianza	276.2832	0.1171	0.3441	0.0213	174.1826	0.0026	0.0001	68.8654
Desviación Estándar	16.6218	0.3423	0.5866	0.1459	13.1978	0.0508	0.008	8.2985
Simetría	-0.7022	0.9253	0.4095	0.3174	-0.9454	-0.9931	0.9551	0.2203

Tabla A.9 Estadísticos modelo dos - clase uno

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	59.37	2.01	7.347	2.031	104.376	0.352	0.0299	-34.4813
1er Cuartil	62.1967	2.0815	7.7177	2.1165	110.208	0.3702	0.0327	-2.1197
Mediana	64.0995	2.166	8.0005	2.1965	115.1585	0.381	0.0337	0.1664
Media	63.9552	2.1624	7.9801	2.1874	114.161	0.3805	0.0339	-1.091
3er Cuartiles	65.5222	2.216	8.1798	2.2355	116.7517	0.3902	0.0353	1.7552
Máximo	68.986	2.364	8.665	2.336	122.189	0.411	0.0382	6.3126
Rango	9.616	0.354	1.318	0.305	17.813	0.059	0.0083	40.7939
Rangos Inter cuartil	3.3255	0.1345	0.462	0.119	6.5437	0.02	0.0026	3.8749
Varianza	5.9707	0.0082	0.1019	0.0075	19.9625	0.0003	0	50.6524
Desviación Estándar	2.4435	0.0906	0.3193	0.0868	4.4679	0.016	0.002	7.117
Simetría	-0.1091	0.4036	0.1446	-0.0001	-0.2151	0.1204	-0.07	-3.5317

Tabla A.10. Estadísticos modelo dos – clase dos

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	94.077	1.37	7.244	1.779	132.11	0.414	0.0126	-13.7951
1er Cuartil	100.6403	1.4508	8.034	1.9033	140.734	0.4408	0.014	-4.2358
Mediana	103.4135	1.492	8.455	1.9625	144.3185	0.454	0.0144	-0.1371
Media	103.1344	1.4931	8.4736	1.9674	144.7818	0.4539	0.0145	0.7069
3er Cuartiles	105.7635	1.5265	8.9597	2.0217	150.2955	0.468	0.015	4.0107
Máximo	113.168	1.601	9.944	2.362	157.212	0.494	0.0164	36.6439
Rango	19.091	0.231	2.7	0.583	25.102	0.08	0.0038	50.439
Rangos Intercuartil	5.1232	0.0757	0.9258	0.1185	9.5615	0.0272	0.001	8.2465
Varianza	16.6248	0.0029	0.3957	0.0097	35.5072	0.0004	0	58.4778
Desviación Estándar	4.0774	0.0543	0.6291	0.0984	5.9588	0.0199	0.0008	7.6471
Simetría	-0.0714	0.0016	0.0154	0.7604	0.1069	-0.2102	-0.1086	1.6696

Tabla A.11. Estadísticos modelo dos – clase tres

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	72.387	1.351	7.217	1.7	127.873	0.252	0.0144	-26.7928
1er Cuartil	76.941	1.545	7.9262	1.8425	136.7847	0.3383	0.0159	-5.4758
Mediana	78.8085	2.0385	8.197	2.028	140.2175	0.356	0.025	0.3395
Media	83.1122	1.9725	8.1614	2.0217	139.6608	0.3689	0.0225	-1.0798
3er Cuartiles	89.3478	2.2838	8.3948	2.192	143.652	0.4255	0.0262	3.8822
Máximo	104.36	2.941	9.637	2.32	150.021	0.473	0.0273	25.6509
Rango	31.973	1.59	2.42	0.62	22.148	0.221	0.0129	52.4437
Rangos Intercuartil	12.4068	0.7388	0.4685	0.3495	6.8673	0.0872	0.0103	9.3581
Varianza	85.0104	0.1565	0.2328	0.0345	33.2177	0.004	0	119.3787
Desviación Estándar	9.2201	0.3956	0.4825	0.1859	5.7635	0.0629	0.0049	10.9261
Simetría	1.0642	0.026	0.5018	-0.0146	-0.4054	0.0806	-0.7731	-0.1549

- Histogramas por modelo y por modelo-clase

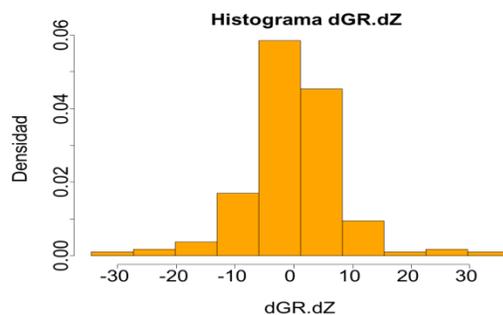
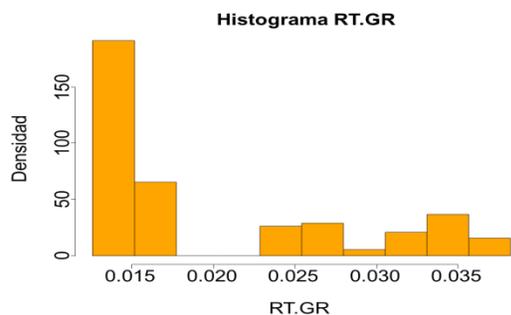
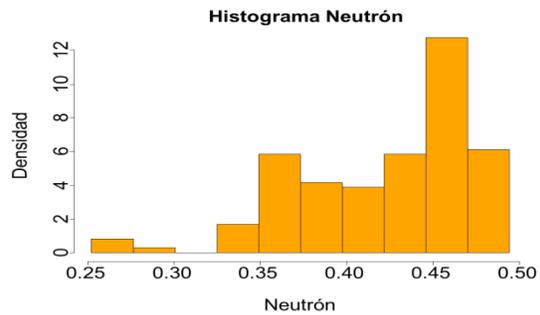
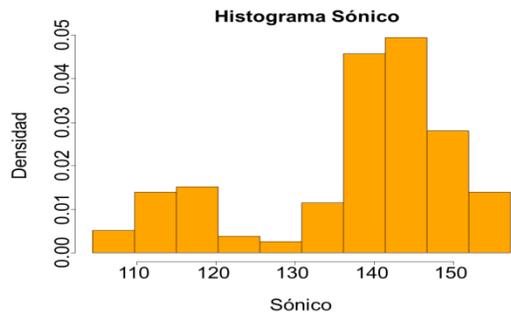
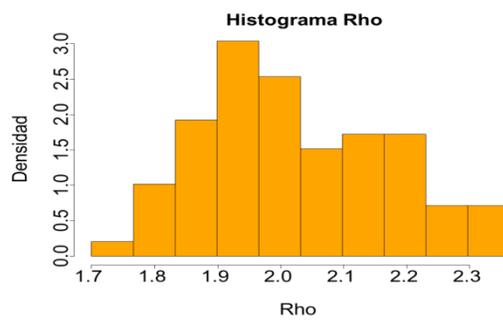
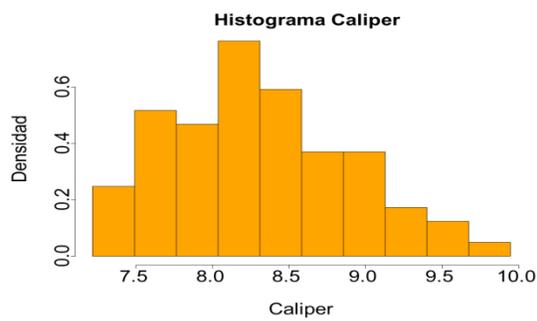
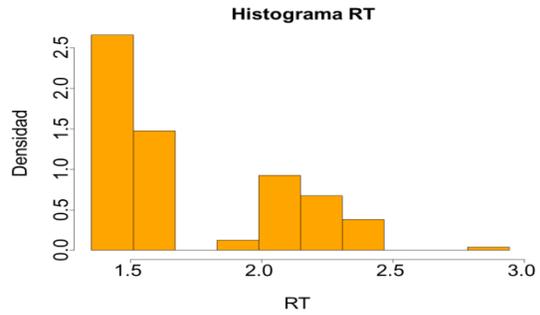
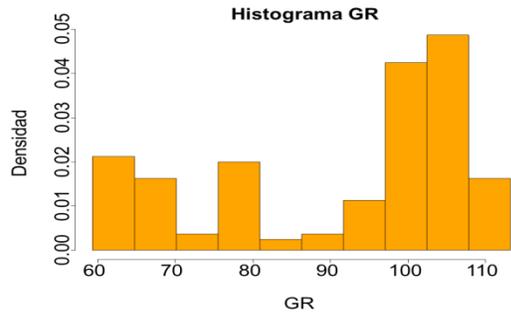


Figura A.16. Histogramas del modelo dos

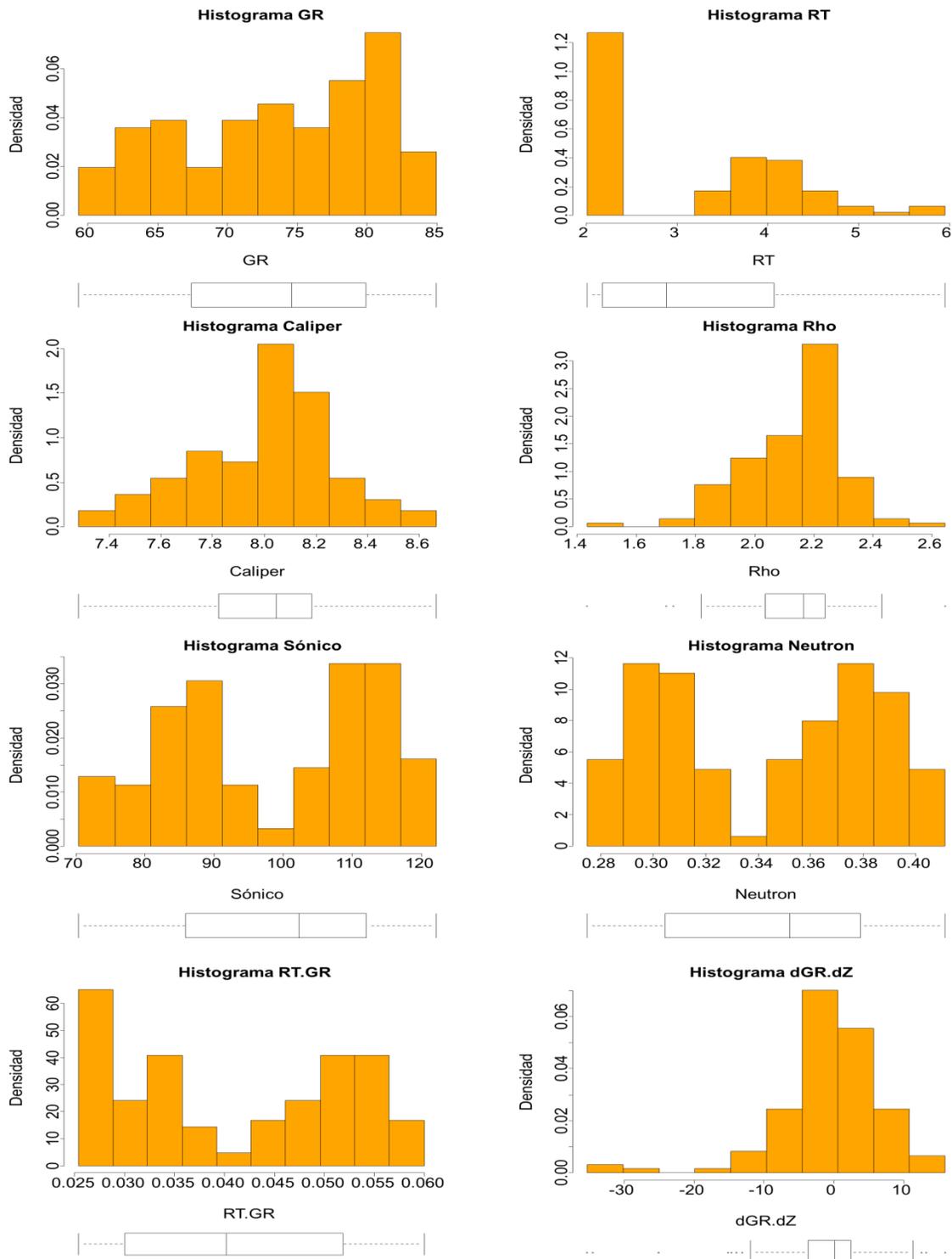


Figura A.17. Histogramas del modelo dos, clase uno

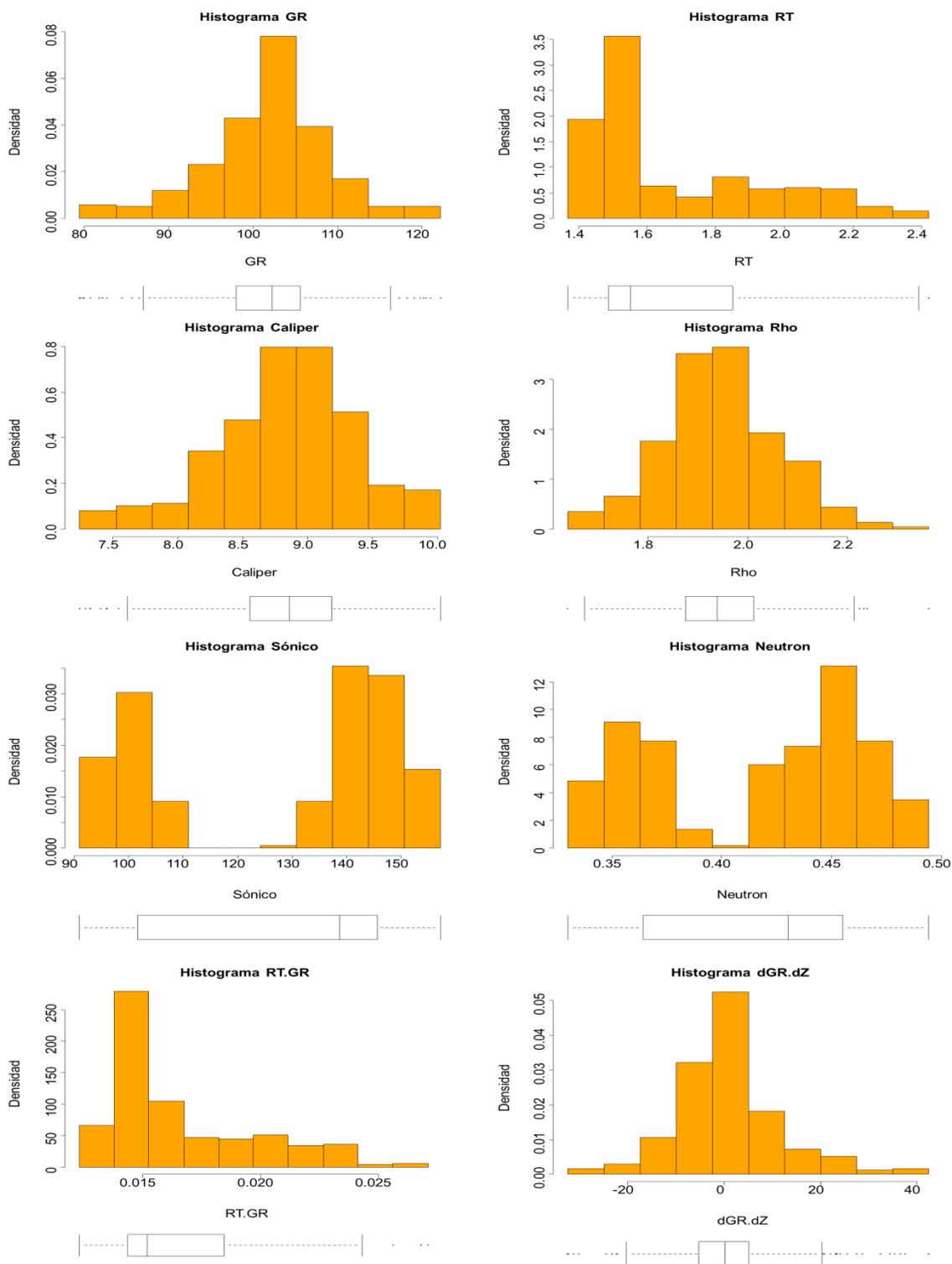


Figura A.18. Histogramas del modelo dos, clase dos

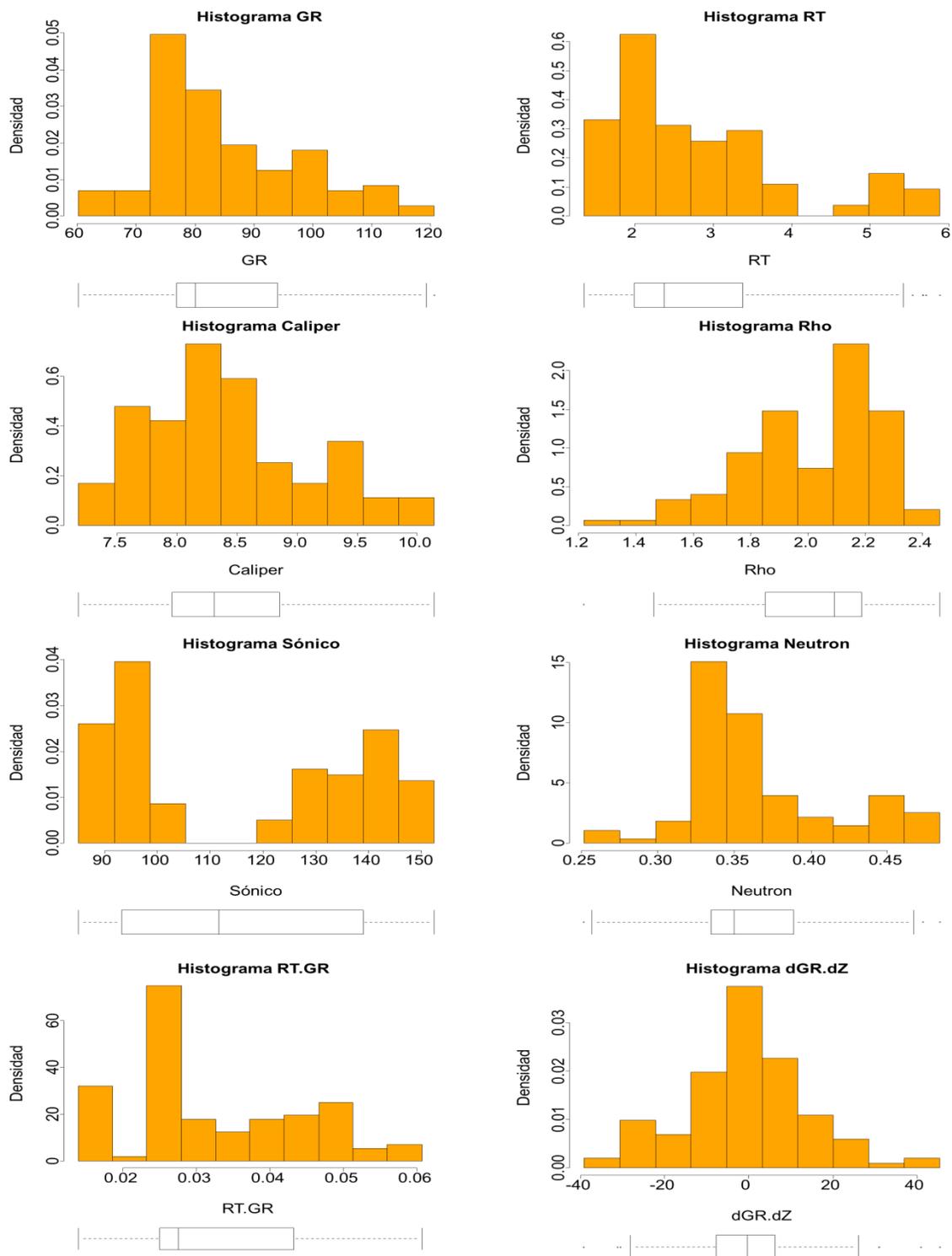


Figura A.19. Histogramas del modelo dos, clase 3

- Gráficos cuantil-cuantil, modelo dos

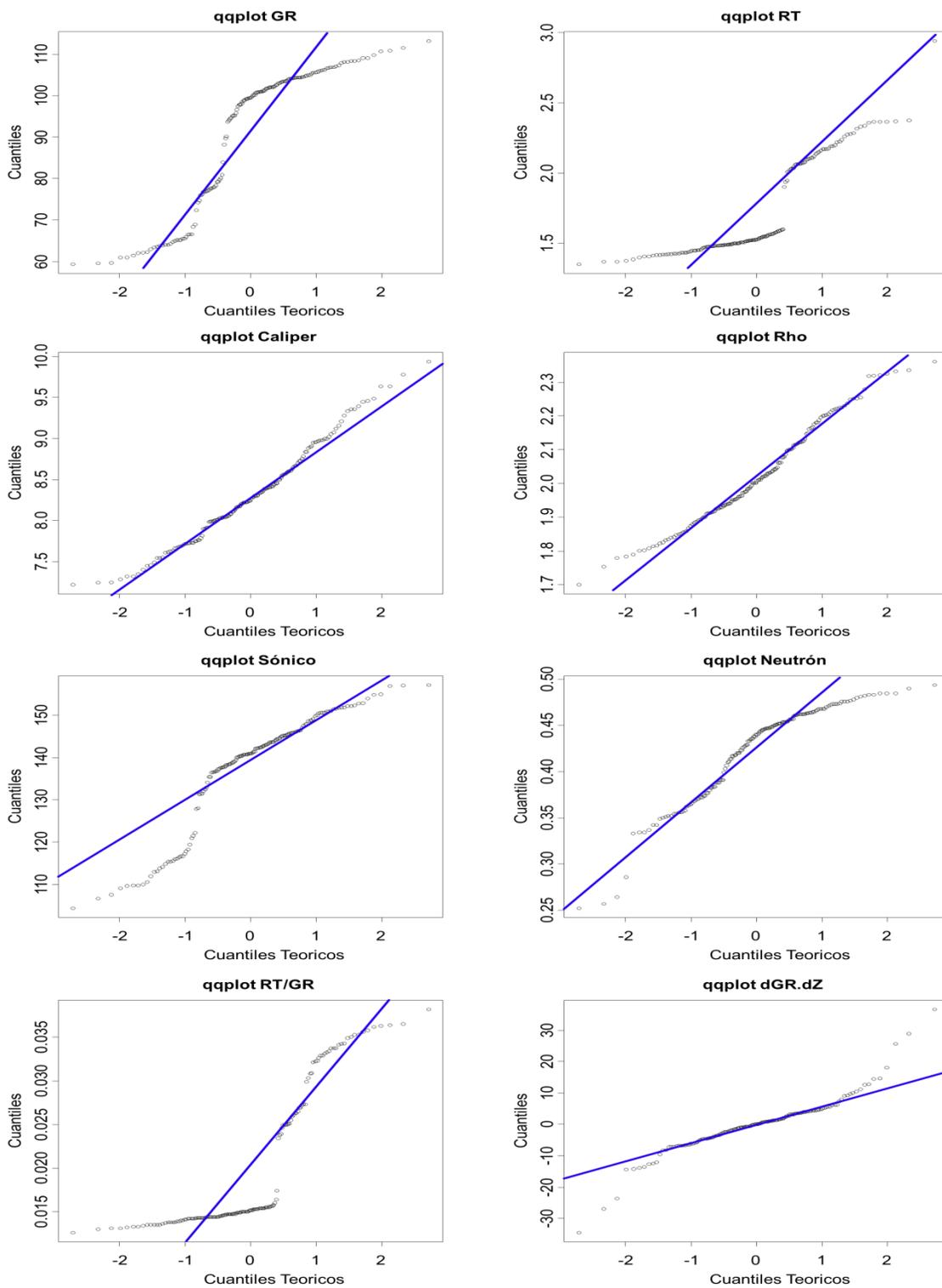


Figura A.20. Gráficos cuantil-cuantil, modelo dos

- Gráfico de dispersión, modelo dos

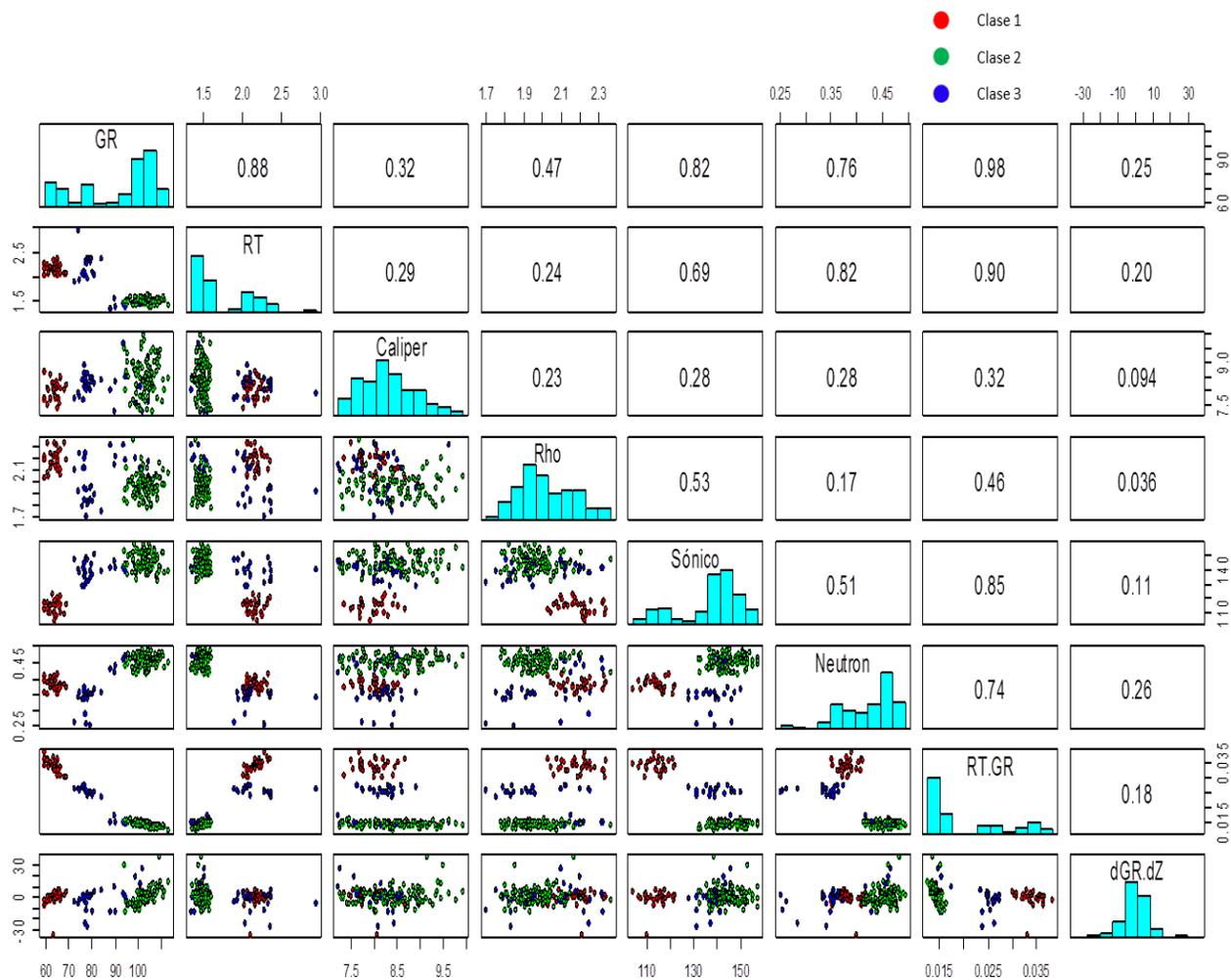


Figura A.21. Gráfico de dispersión modelo dos

C. modelo tres

- Estadísticos

Tabla A.12. Estadísticos modelo tres

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	60.234	1.513	7.283	1.218	85.102	0.316	0.0142	-39.12
1er Cuartil	81.2452	1.91	8.3505	1.857	95.0772	0.341	0.0194	-8.72
Mediana	93.273	2.143	8.909	1.991	98.8815	0.356	0.0232	-0.42
Media	93.0274	2.4784	8.8235	1.988	99.5198	0.3558	0.0273	0.0157
3er Cuartiles	102.7597	2.5993	9.3305	2.1547	103.2622	0.368	0.0301	8.11
Máximo	122.157	5.886	10.142	2.46	118.483	0.408	0.0607	45.31
Rango	61.923	4.373	2.859	1.242	33.381	0.092	0.0465	84.43
Rangos Intercuartil	21.5145	0.6892	0.98	0.2977	8.185	0.027	0.0107	16.83
Varianza	191.3098	0.8982	0.4817	0.0423	40.6209	0.0004	0.0001	219.8527
Desviación Estándar	13.8315	0.9477	0.694	0.2056	6.3735	0.0188	0.011	14.8274
Simetría	0.066	2.0072	-0.3996	-0.4018	0.4727	0.1765	1.1739	0.2112

Tabla A.13. Estadísticos modelo tres – clase uno

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	74.312	2.089	7.283	2.011	102.214	0.351	0.0254	-25.09
1er Cuartil	78.7007	2.169	7.6498	2.097	107.0385	0.367	0.0268	-2.9475
Mediana	79.9345	2.1985	7.94	2.1995	109.3995	0.378	0.0274	-0.48
Media	79.9287	2.1979	7.9164	2.1804	109.7886	0.3789	0.0275	-0.832
3er Cuartiles	81.2717	2.222	8.1225	2.2445	113.167	0.391	0.0281	2.3125
Máximo	84.494	2.367	8.632	2.348	118.483	0.408	0.0301	9.18
Rango	10.182	0.278	1.349	0.337	16.269	0.057	0.0047	34.27
Rangos Intercuartil	2.571	0.053	0.4727	0.1475	6.1285	0.024	0.0013	5.26
Varianza	5.983	0.0038	0.1352	0.0081	19.5168	0.0002	0	38.0437
Desviación Estándar	2.446	0.0616	0.3677	0.0903	4.4178	0.0151	0.0011	6.168
Simetría	-0.388	0.8054	0.2062	-0.3557	0.1788	0.0496	0.7711	-1.9061

Tabla A.14. Estadísticos modelo tres – clase dos

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	79.501	1.513	8.54	1.64	91.001	0.33	0.0142	-32.44
1er Cuartil	92.248	1.8252	8.8505	1.8508	97.1225	0.35	0.0179	-10.33
Mediana	98.485	1.9405	9.1185	1.9285	99.908	0.357	0.0199	1.73
Media	99.6596	1.964	9.1669	1.9404	99.8346	0.3584	0.0199	0.4855
3er Cuartiles	106.1963	2.1082	9.4443	2.0275	103.1045	0.368	0.0218	10.865
Máximo	122.157	2.42	10.027	2.239	108.686	0.392	0.0271	42.44
Rango	42.656	0.907	1.487	0.599	17.685	0.062	0.0129	74.88
Rangos Intercuartil	13.9483	0.283	0.5938	0.1767	5.982	0.018	0.0039	21.195
Varianza	105.7337	0.0392	0.1442	0.0204	15.6101	0.0002	0	228.1822
Desviación Estándar	10.2827	0.198	0.3797	0.1427	3.951	0.0138	0.0028	15.1057
Simetría	0.1618	0.0661	0.429	0.0974	-0.025	0.0067	0.2583	0.1766

Tabla A.15. Estadísticos modelo tres – clase tres

	GR	RT	Caliper	Rho	Sónico	Neutron	RT/GR	dGR/dZ
Mínimo	60.234	1.868	7.369	1.218	85.102	0.316	0.021	-39.12
1er Cuartil	75.1167	2.8562	7.8972	1.751	91.0562	0.3288	0.0345	-10.7925
Mediana	84.3635	3.3755	8.4445	2.087	93.244	0.338	0.0433	-0.925
Media	86.3124	3.6477	8.5901	1.9869	93.7558	0.339	0.0421	-0.4923
3er Cuartiles	96.8085	4.1365	9.358	2.219	96.9825	0.3482	0.0494	10.455
Máximo	120.628	5.886	10.142	2.46	102.272	0.369	0.0607	45.31
Rango	60.394	4.018	2.773	1.242	17.17	0.053	0.0397	84.43
Rangos Intercuartil	21.6918	1.2803	1.4608	0.468	5.9263	0.0195	0.0149	21.2475
Varianza	238.9303	1.1322	0.6353	0.0812	15.2103	0.0002	0.0001	298.937
Desviación Estándar	15.4574	1.0641	0.7971	0.2849	3.9	0.0134	0.009	17.2898
Simetría	0.4942	0.7132	0.2226	-0.5392	0.1699	0.3456	-0.1784	0.2509

- Histogramas globales y por clase.

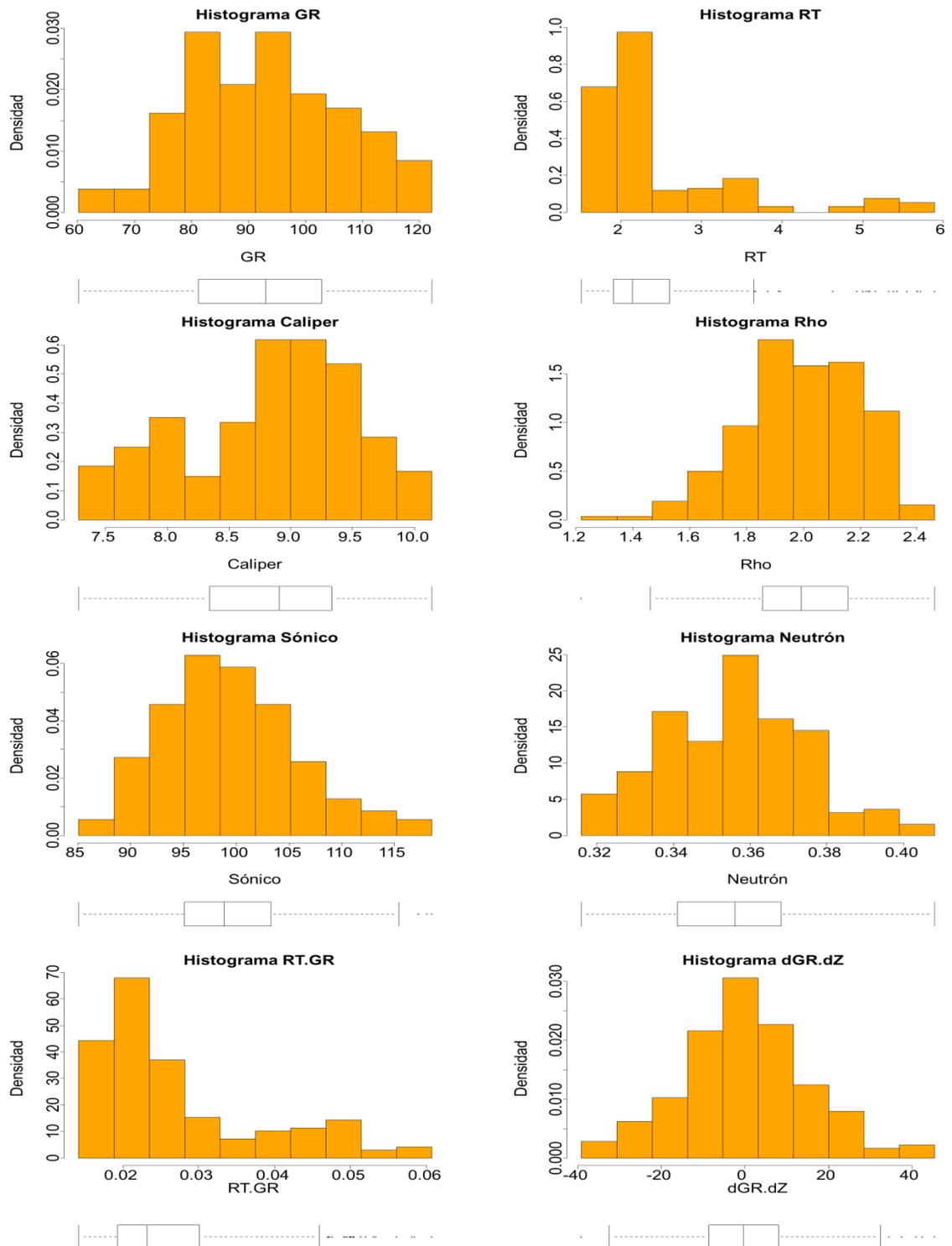


Figura A.22 Histogramas del modelo tres

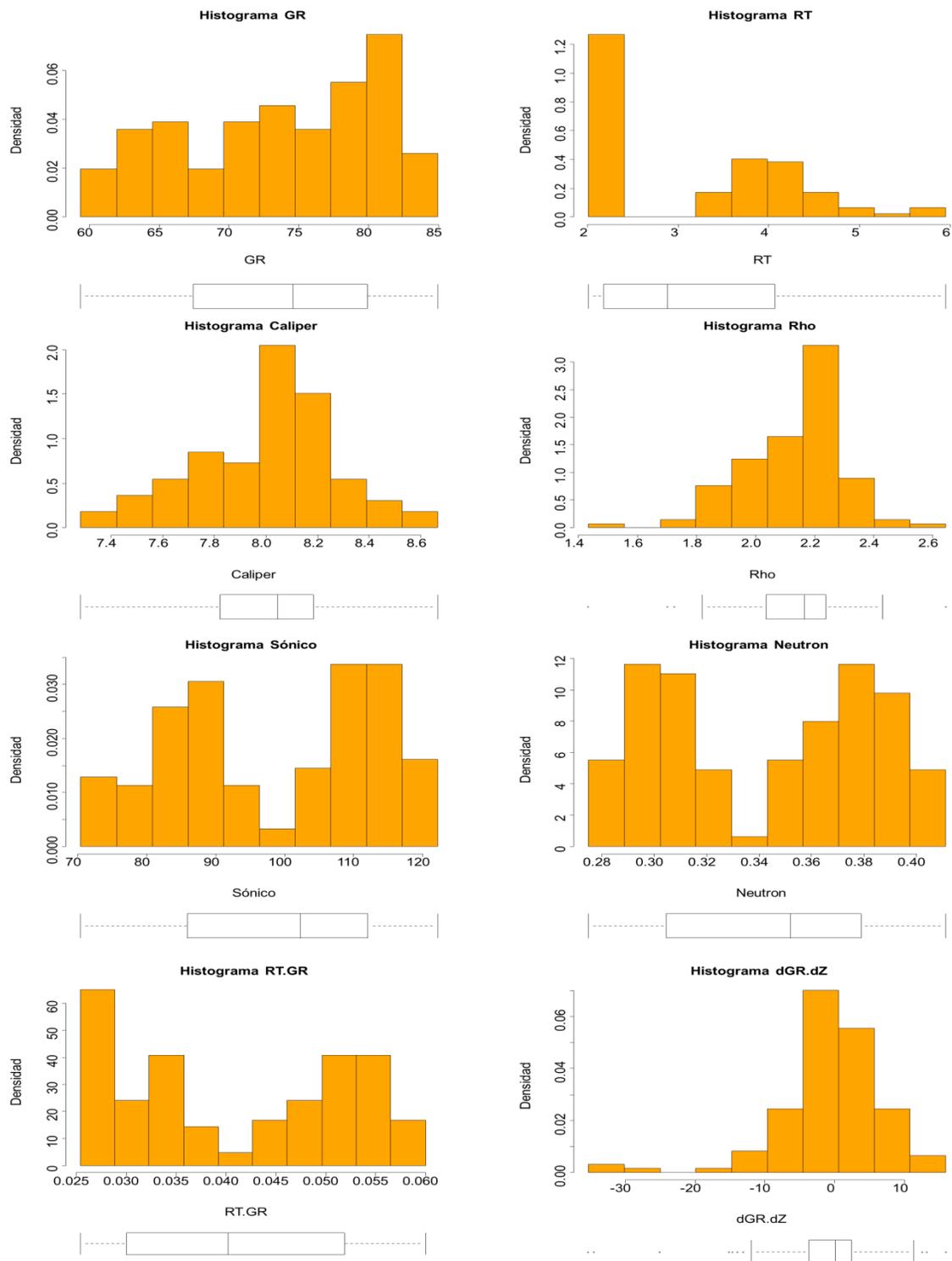


Figura A.23. Histogramas del modelo tres, clase uno

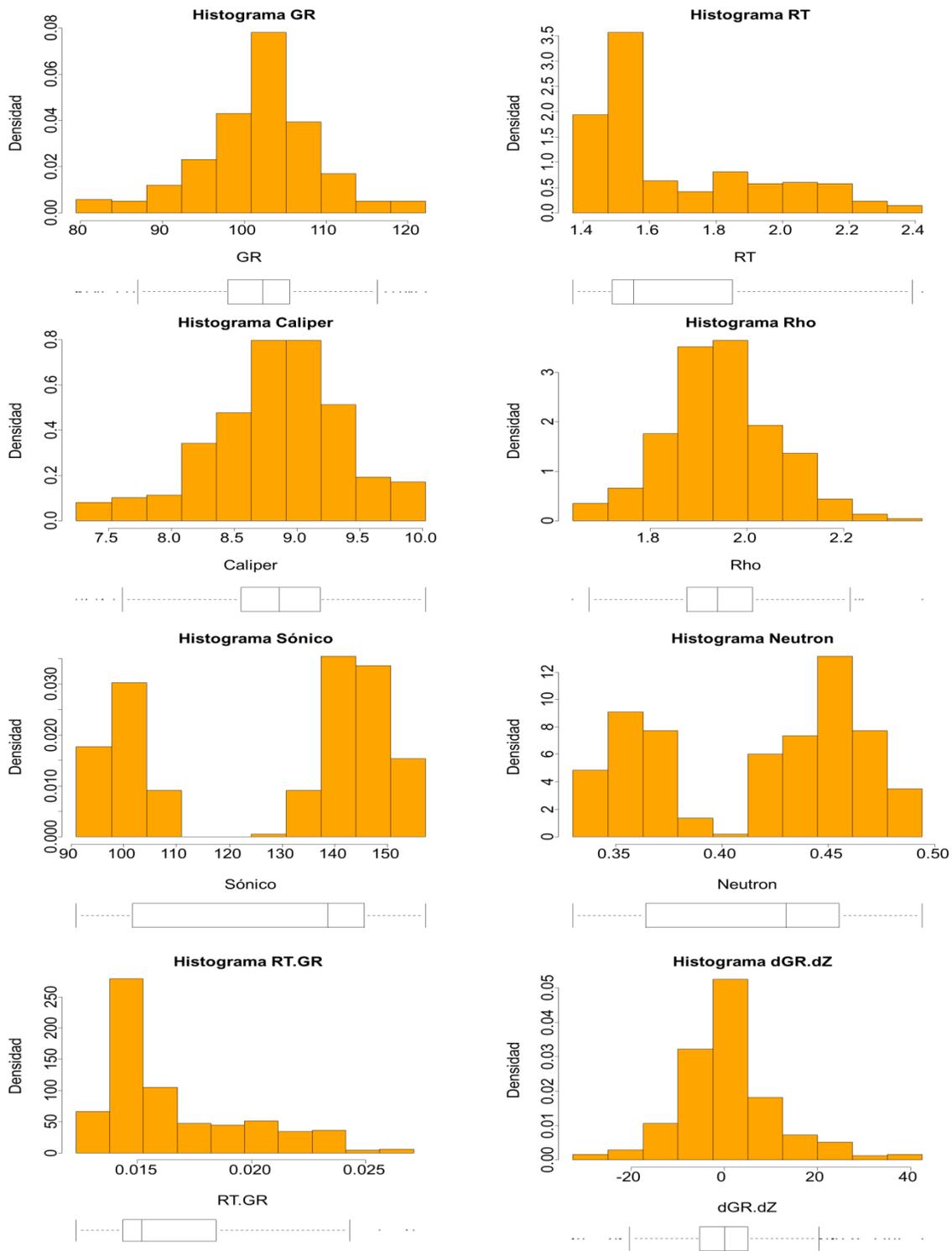


Figura A.24. Histogramas del modelo tres, clase dos

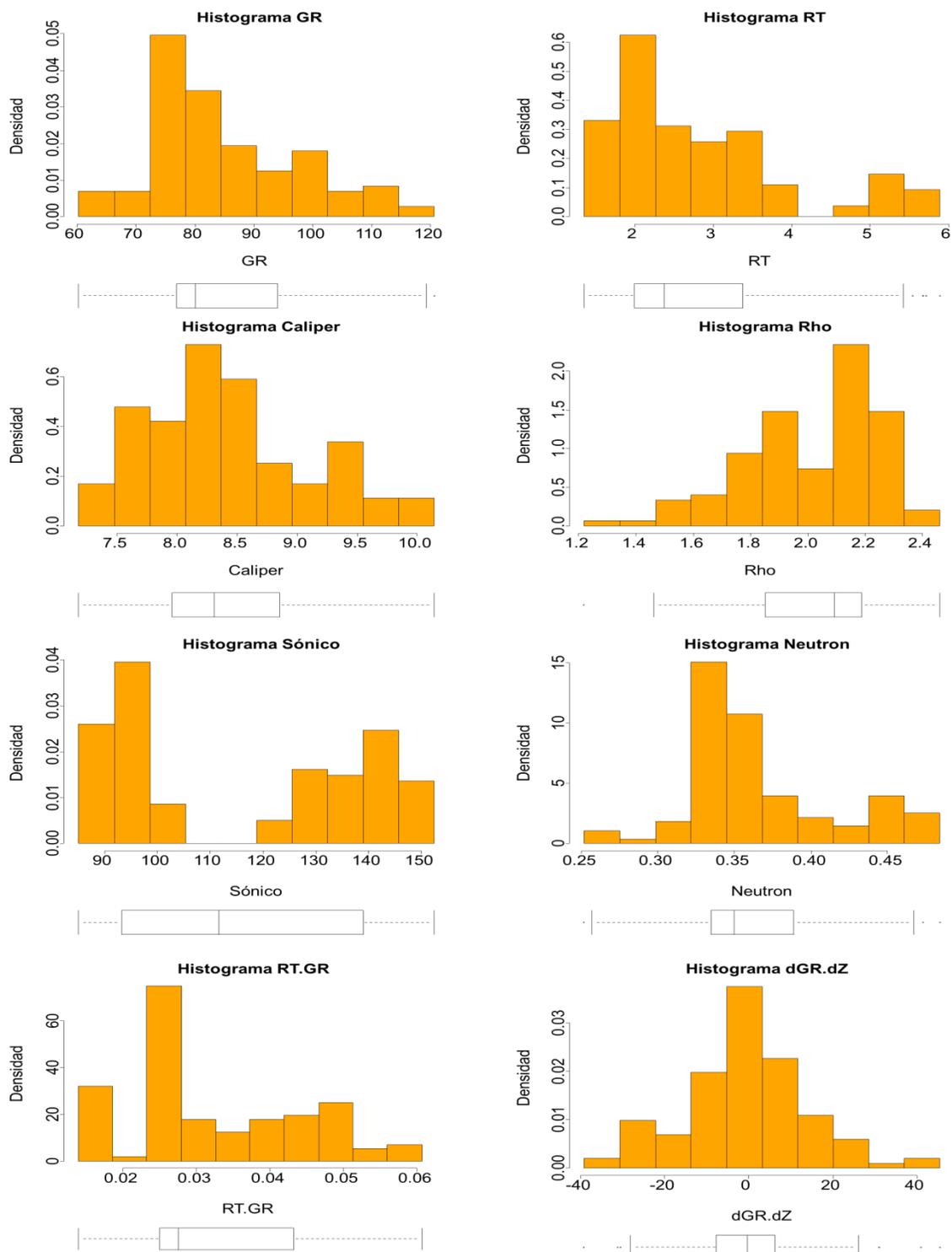


Figura A.25. Histogramas del modelo tres, clase 3

- Gráficos cuantil-cuantil, modelo tres

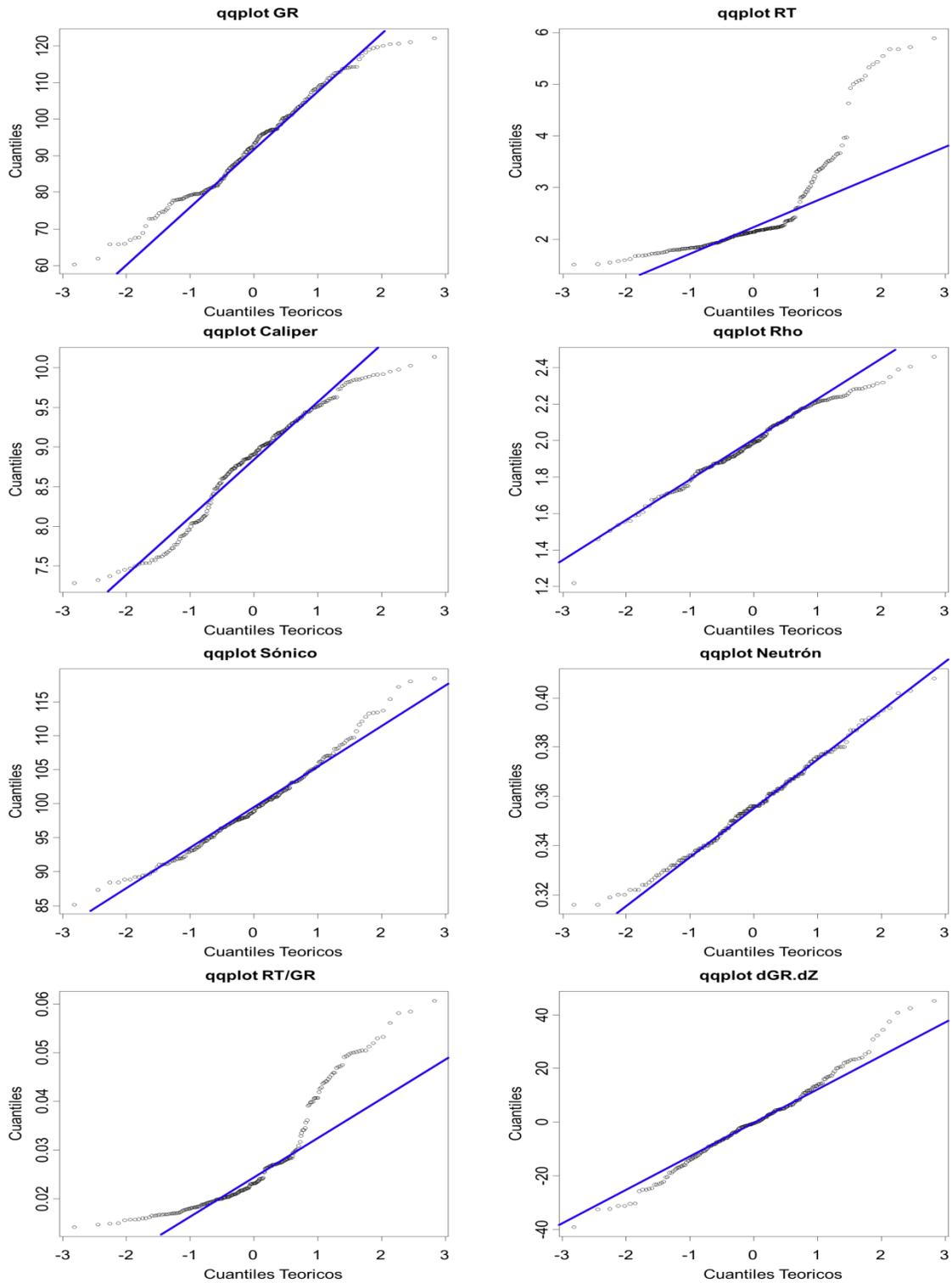


Figura A.26. Gráficos cuantil- cuantil, modelo tres

- Gráfico de dispersión, modelo tres

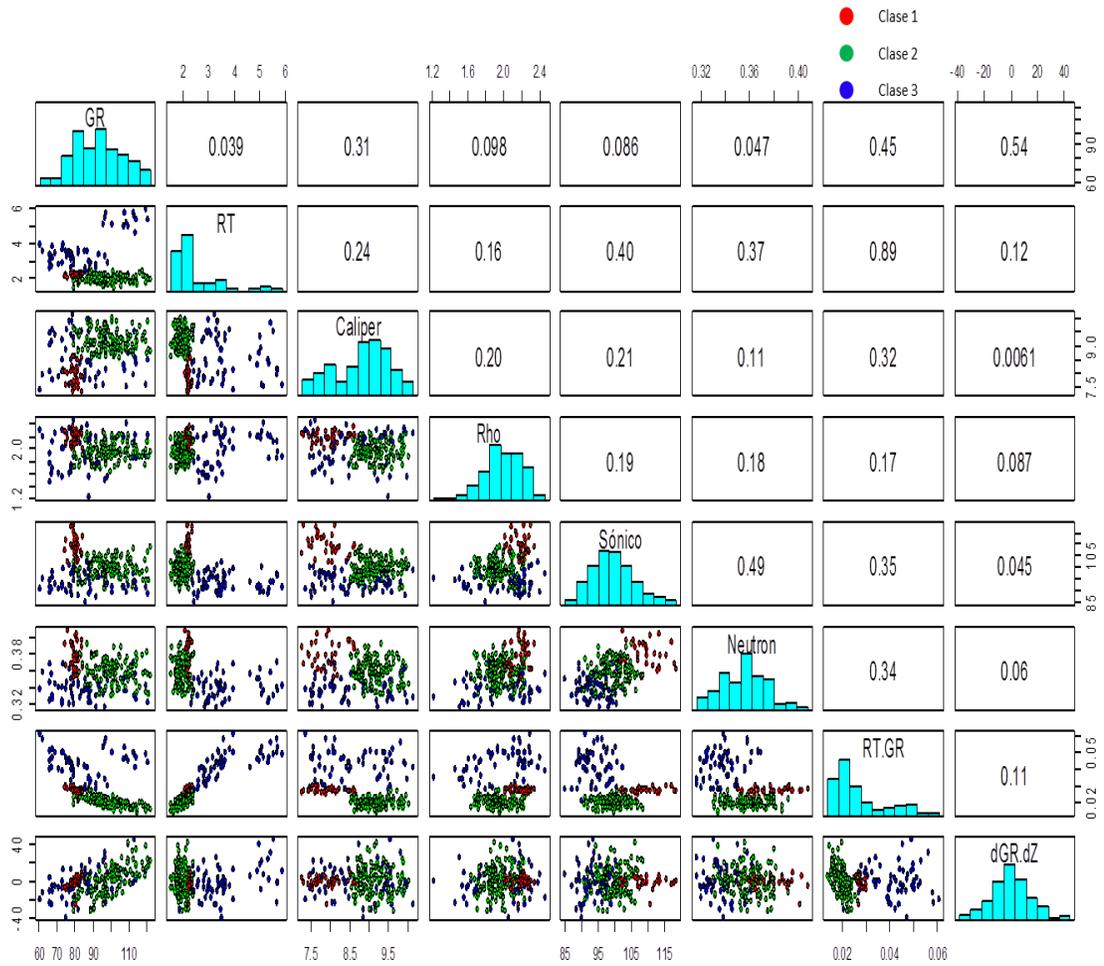


Figura A.27. Gráfico de dispersión, Gráficos cuantil modelo tres

Apéndice B

Análisis Exploratorio de Datos (AED) Registros Reales

III. Análisis por Litofacie (Pozos 311, 312, 318, 331, 332, 333, 351 y 353)

a) Estadísticos por clase.

Tabla B.1 Estadísticos litofacie uno (Arenisca)

	GR	RT	CALI	DEN	DT	NPHI	RT/GR
Minimos	30.753	3.508	3.33	2.2248	58.483	0.1	0.0725
Cuantiles1	42.276	8.1367	6.773	2.503	69.2451	0.1917	0.1761
Medianas	45.443	11.1945	6.836	2.5394	72.8475	0.218	0.248
Medias	44.6331	12.5545	6.864	2.5313	72.6204	0.2213	0.293
Cuantiles3	47.8492	14.787	6.919	2.5699	76.0893	0.2411	0.3424
Maximo	49.998	58.636	8.583	2.6667	88.604	0.4165	1.843
Rangos	19.245	55.128	5.253	0.4419	30.121	0.3165	1.7705
Rangosint	5.5733	6.6503	0.146	0.0669	6.8441	0.0494	0.1664
Varianzas	16.2418	43.9621	0.1133	0.0035	24.3793	0.0024	0.0353
Desvs	4.0301	6.6304	0.3365	0.0592	4.9375	0.0492	0.1878
Simetrias	-0.9163	2.0058	-2.6133	-1.2831	-0.0743	1.1464	2.5469

Tabla B.2 Estadísticos litofacie dos (Arenisca arcillosa)

	GR	RT	CALI	DEN	DT	NPHI	RT/GR
Minimos	50.0032	2.396	3.381	2.2172	65.056	0.1434	0.0436
Cuantiles1	52.5493	4.618	6.821	2.484	76.1975	0.238	0.084
Medianas	54.583	5.705	6.906	2.507	79.2449	0.258	0.1044
Medias	54.3776	6.4561	6.9495	2.5037	79.1416	0.2575	0.1196
Cuantiles3	56.3408	7.424	7.029	2.53	82.2518	0.2765	0.1396
Maximo	58	25.439	8.598	2.618	91.184	0.4162	0.501
Rangos	7.9968	23.043	5.217	0.4008	26.128	0.2728	0.4574
Rangosint	3.7915	2.806	0.208	0.046	6.0543	0.0385	0.0556
Varianzas	5.1409	7.4089	0.1134	0.002	20.2255	0.001	0.0028
Desvs	2.2674	2.7219	0.3367	0.0445	4.4973	0.0319	0.0527
Simetrias	-0.221	1.9759	-2.1534	-1.2145	-0.1963	0.3266	1.9538

Tabla B.3 Estadísticos litofacie tres (Lutita arenosa)

	GR	RT	CALI	DEN	DT	NPHI	RT/GR
Minimos	58.0028	1.255	3.567	1.9546	68.965	0.1654	0.0195
Cuantiles1	60.6934	3.392	6.906	2.47	81.4796	0.269	0.0521
Medianas	63.56	3.957	7.018	2.494	84.404	0.288	0.0625
Medias	63.6534	4.2537	7.0933	2.4861	84.3735	0.2868	0.0674
Cuantiles3	66.479	4.646	7.2375	2.513	87.283	0.305	0.0749
Maximo	69.994	21.102	9.196	2.594	118.351	0.3877	0.3445
Rangos	11.9912	19.847	5.629	0.6394	49.386	0.2223	0.325
Rangosint	5.7856	1.254	0.3315	0.043	5.8034	0.036	0.0228
Varianzas	11.2547	2.0343	0.1588	0.0022	19.1218	0.0007	0.0006
Desvs	3.3548	1.4263	0.3985	0.0467	4.3729	0.0273	0.0246
Simetrias	0.1171	2.9713	-0.7049	-2.6149	0.511	-0.133	2.9256

Tabla B.4 Estadísticos Litofacie 4 (Lutita)

	GR	RT	CALI	DEN	DT	NPHI	RT/GR
Minimos	70.002	0.625	4.45	2.174	72.654	0.1884	0.0071
Cuantiles1	71.779	2.607	6.949	2.4381	86.454	0.305	0.0343
Medianas	74.038	2.982	7.155	2.482	88.765	0.321	0.0399
Medias	75.1336	3.0411	7.2968	2.4624	89.2283	0.3244	0.0408
Cuantiles3	77.395	3.4443	7.4677	2.505	91.257	0.3395	0.047
Maximo	94.076	7.859	10.086	2.557	112.879	0.574	0.1107
Rangos	24.074	7.234	5.636	0.383	40.225	0.3856	0.1036
Rangosint	5.616	0.8372	0.5187	0.0669	4.803	0.0345	0.0127
Varianzas	18.3716	0.6095	0.3264	0.0041	26.7504	0.0011	0.0001
Desvs	4.2862	0.7807	0.5713	0.0643	5.1721	0.0332	0.0114
Simetrias	1.2243	0.7032	1.5049	-1.5986	1.1707	1.4935	0.6766

b) Histogramas por litofacie

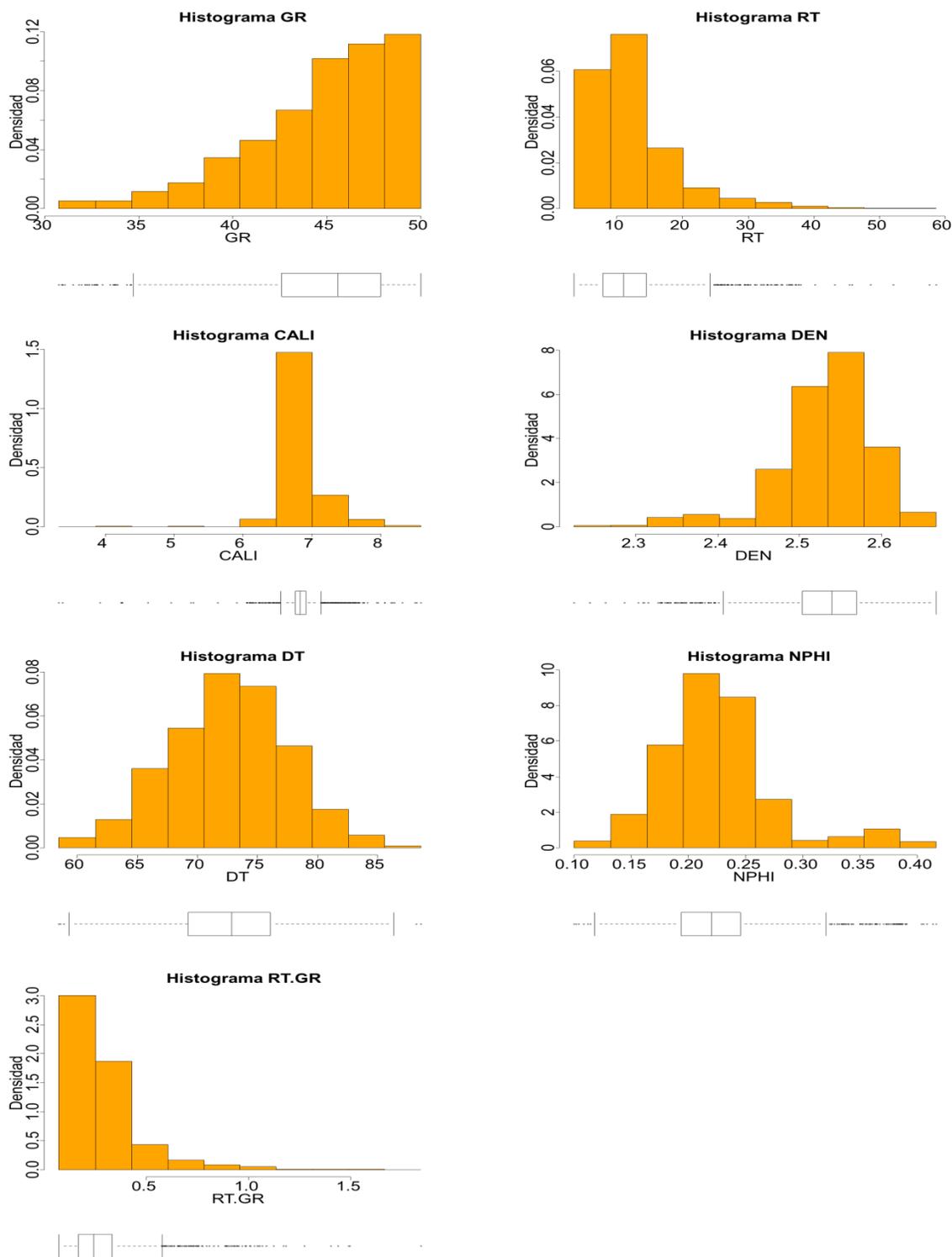


Figura B.1 Histogramas litofacie uno (Areniscas)

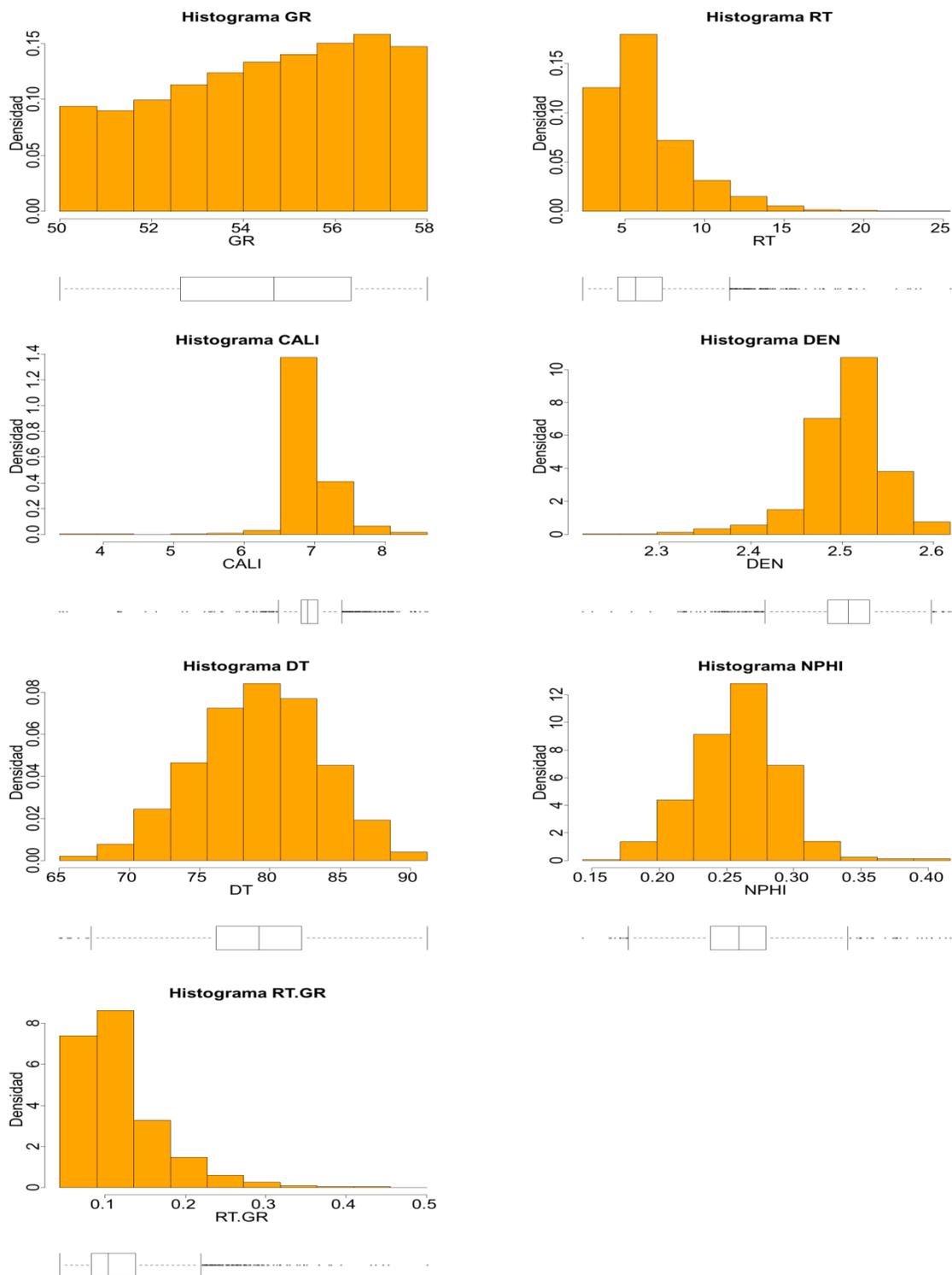


Figura B.2 Histogramas litofacie 2 (Arenisca arcillosa)

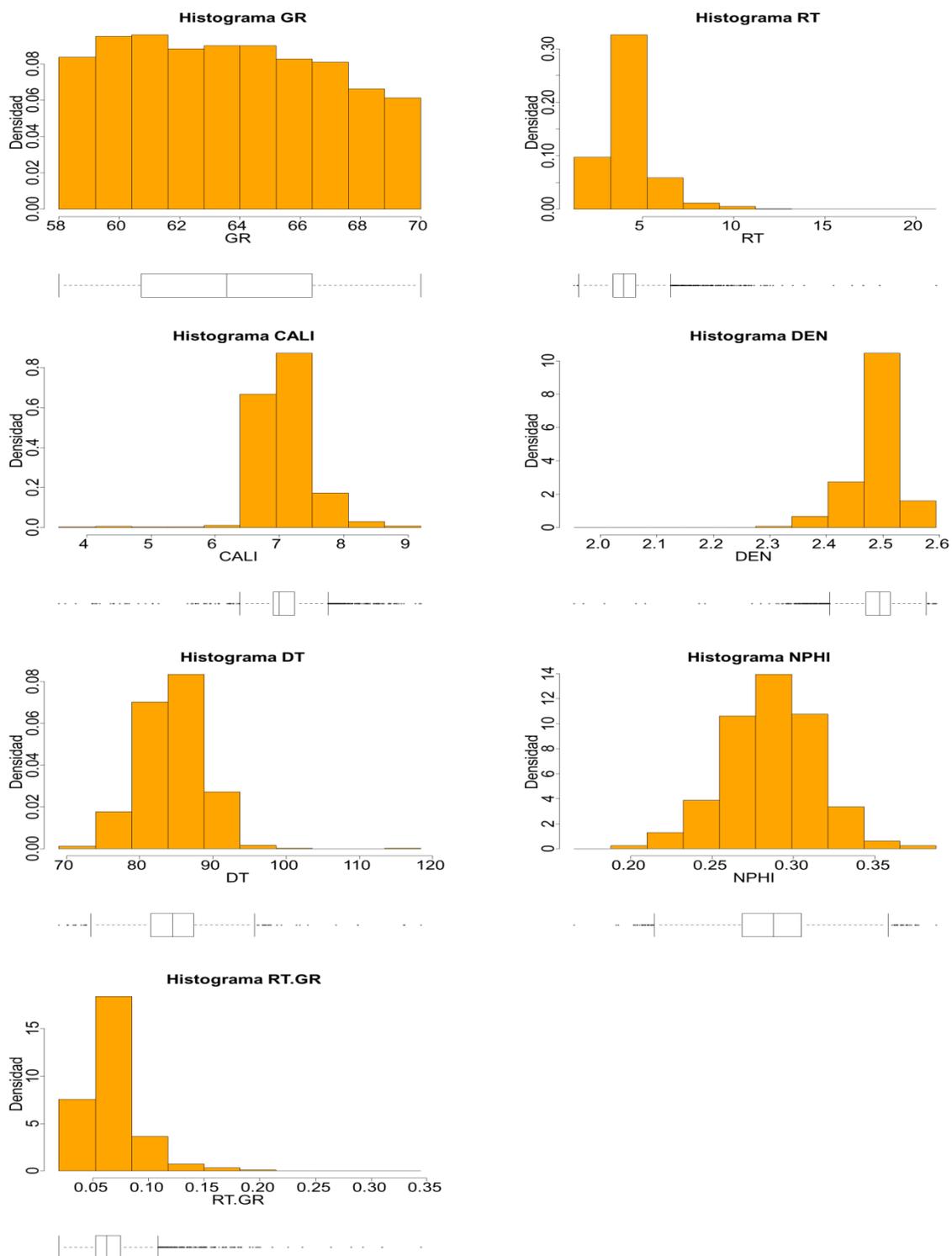


Figura B.3 Histogramas litofacie tres (Lutita – Arenosa)

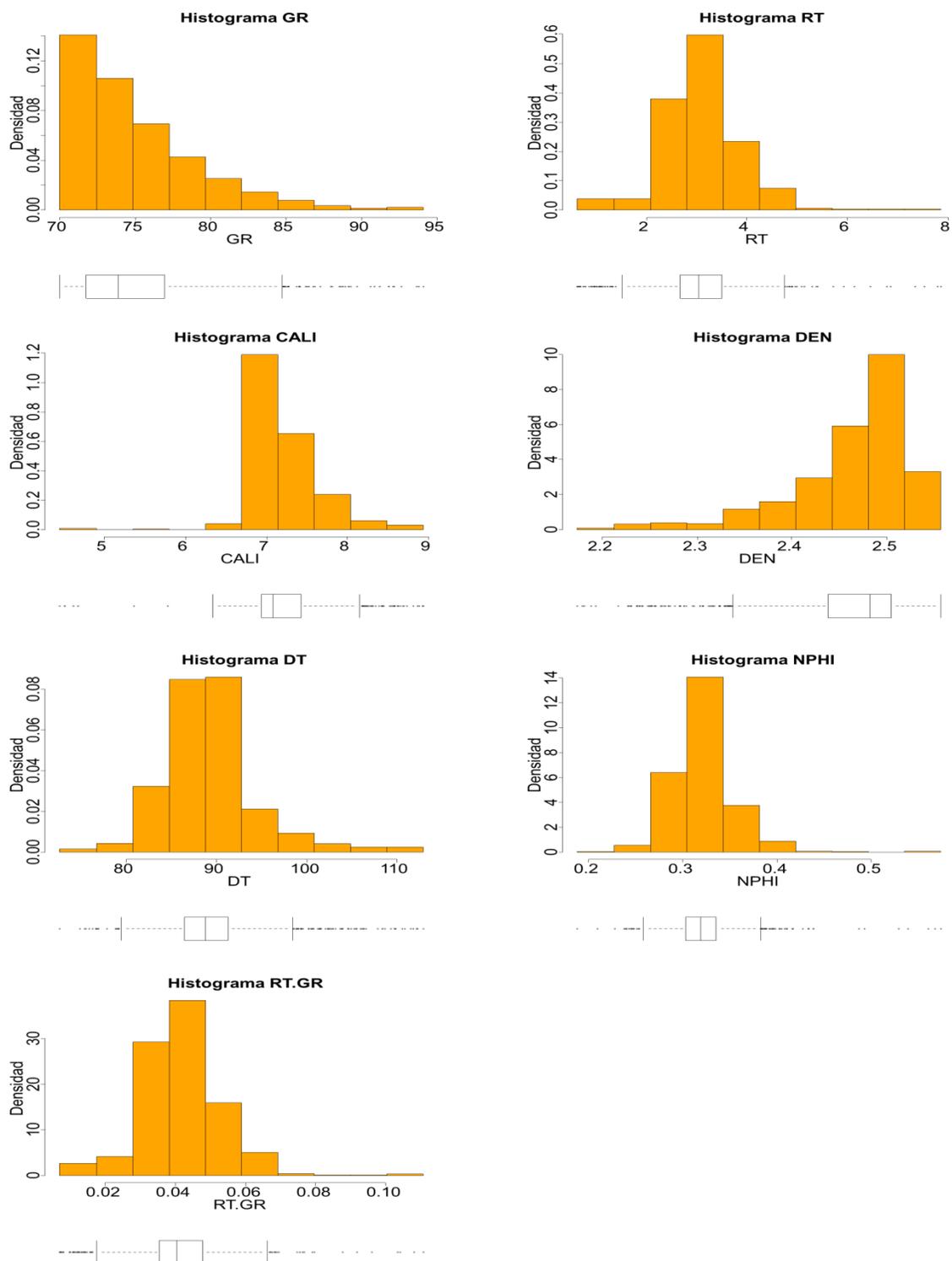


Figura B.4 Histogramas litofacie 4 (Lutita)

c) Gráfico de cajaGráficos de cajaGráficos de caja por litofacie

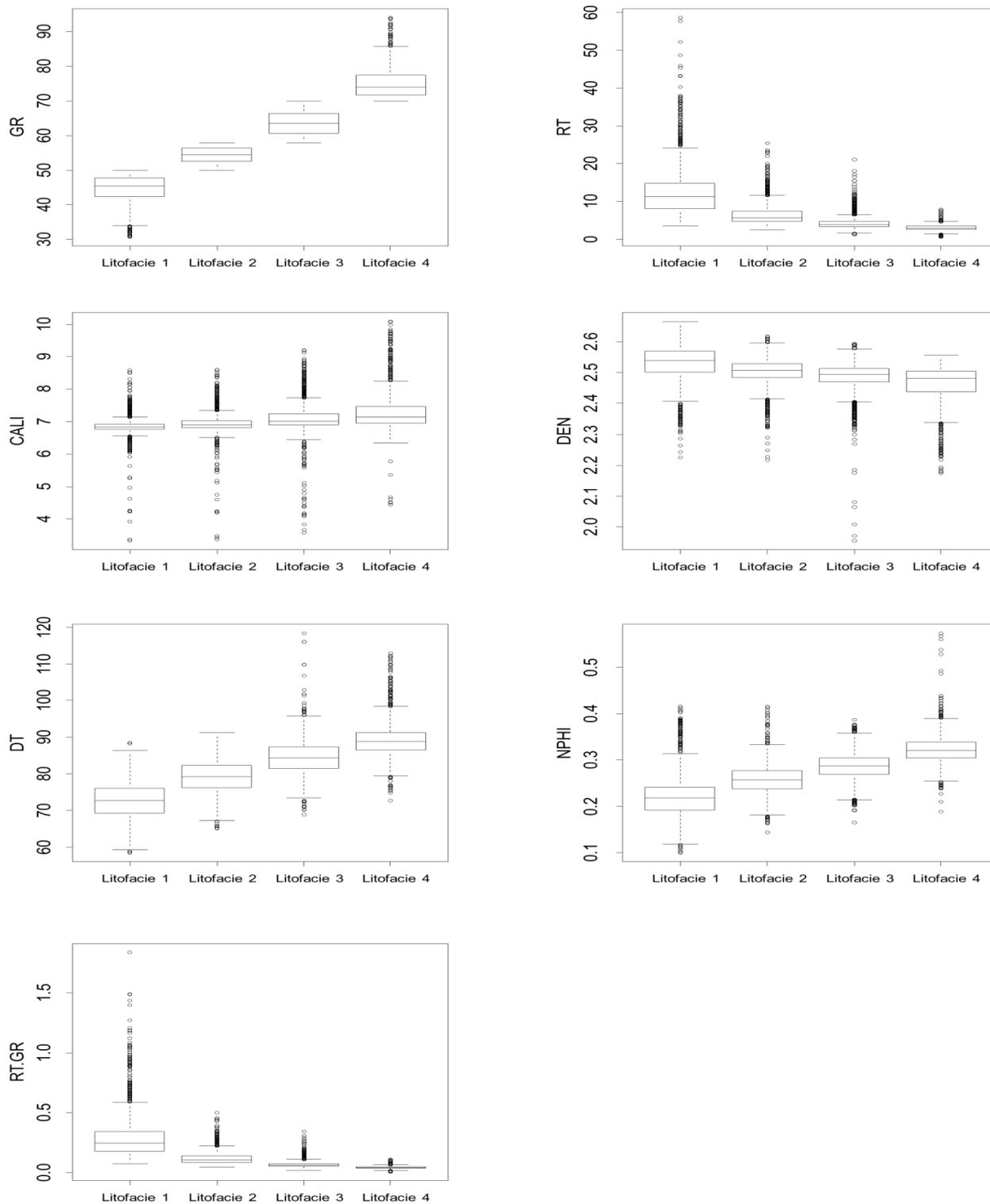


Figura B.5. Gráfico de cajaGráficos de cajaGráficos de caja por Litofacies

d) Gráficos cuantil - cuantil por clase

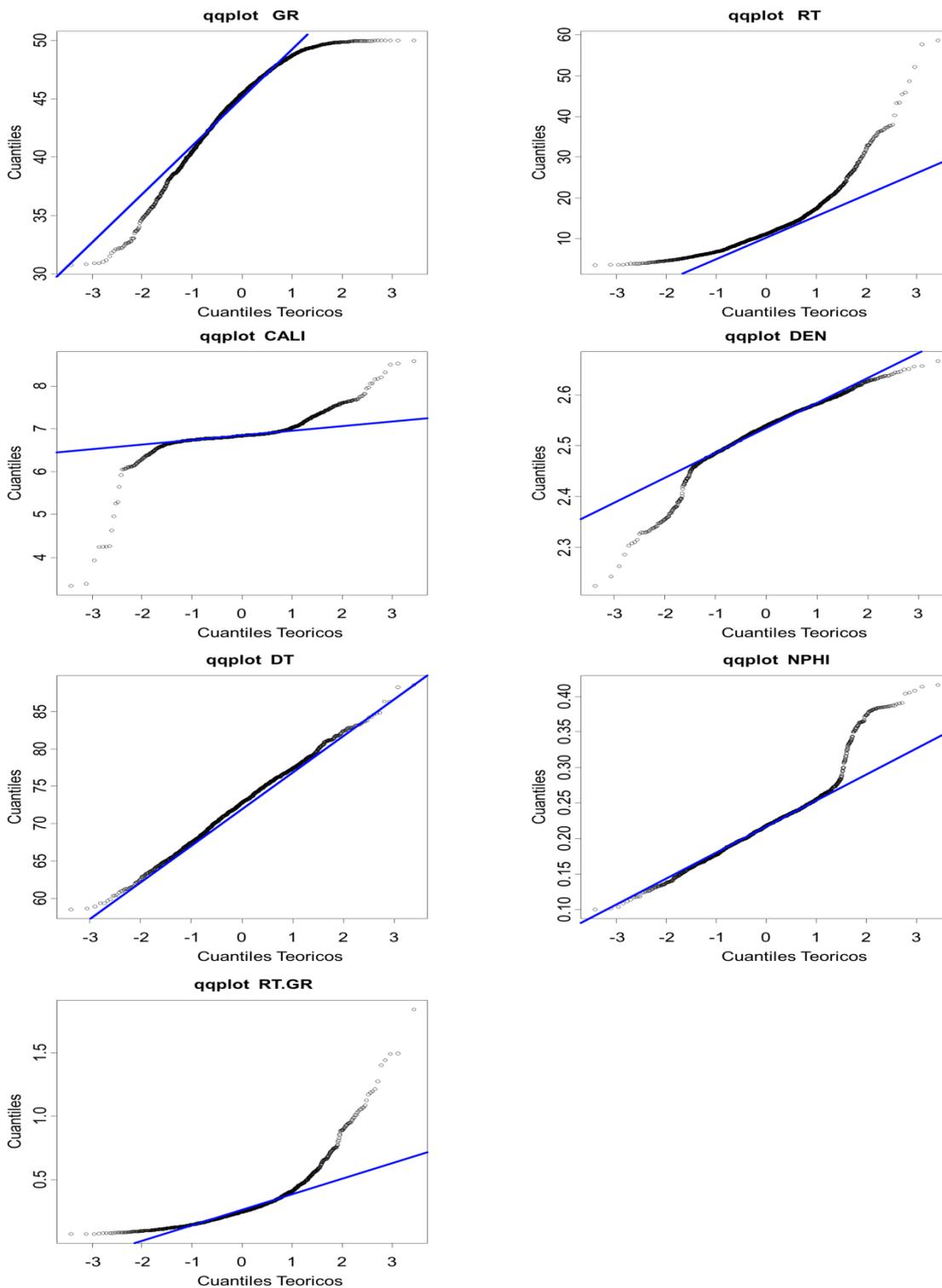


Figura B.6 Gráficos cuantil- cuantil. litofacie uno (Arenisca)

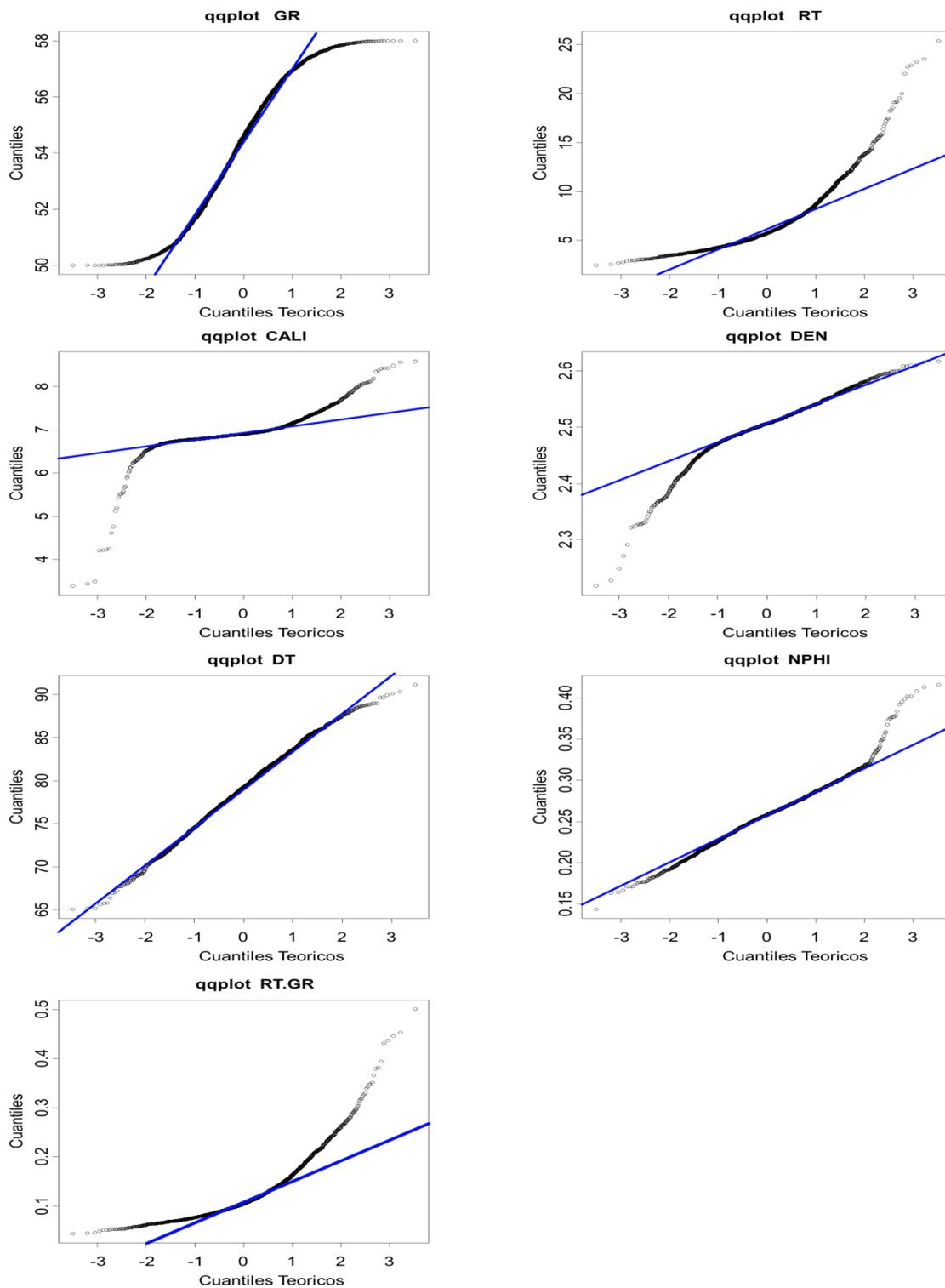


Figura B.7 Gráficos cuantil- cuantil. litofacie dos (Arena Arcillosa)

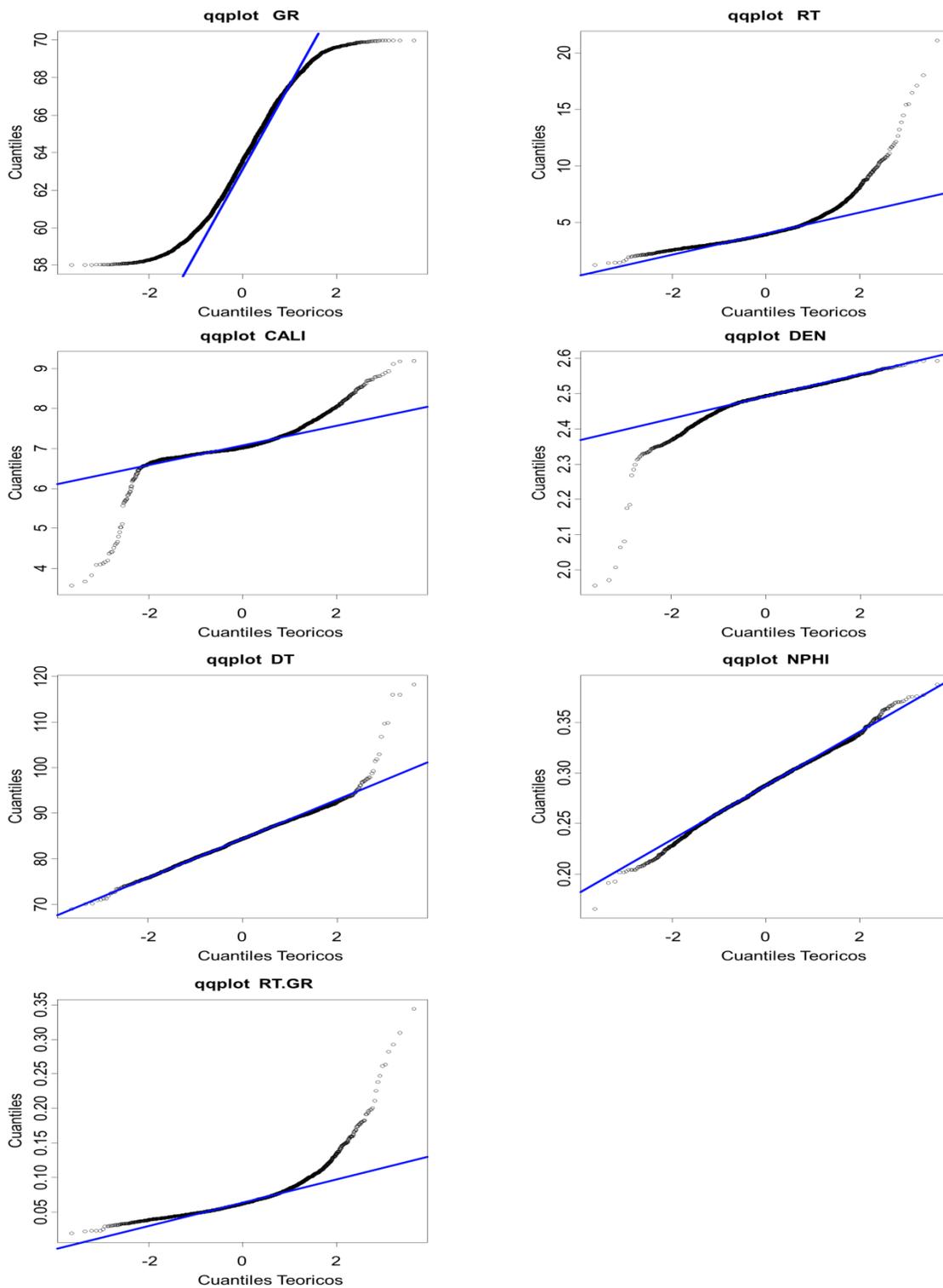


Figura B.8 Gráficos cuantil- cuantil. litofacie tres (Lutita Arenosa)

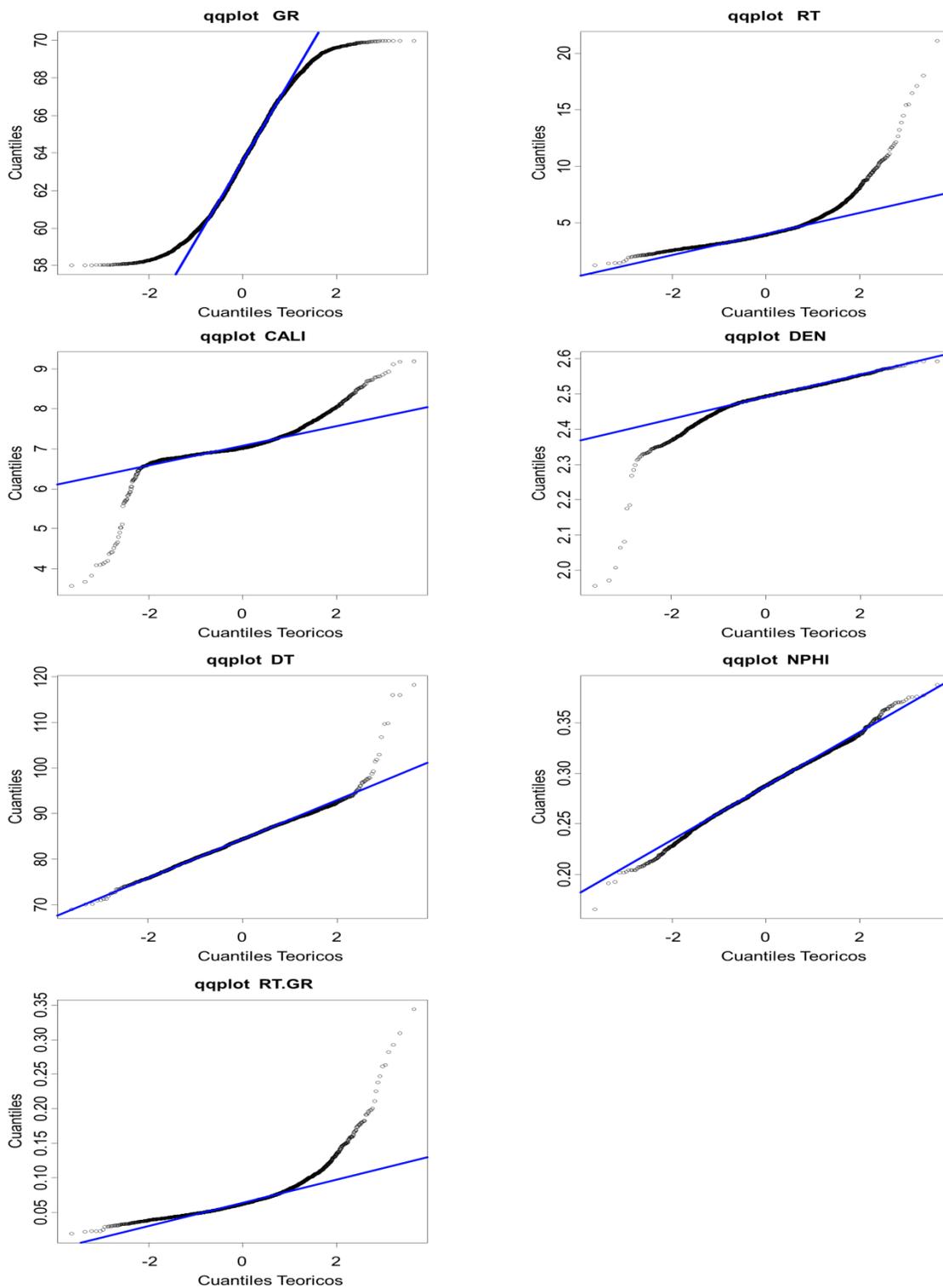


Figura B.9 Gráficos cuantil- cuantil. Litofacie 4 (Lutita)

e) Gráficos de dispersión

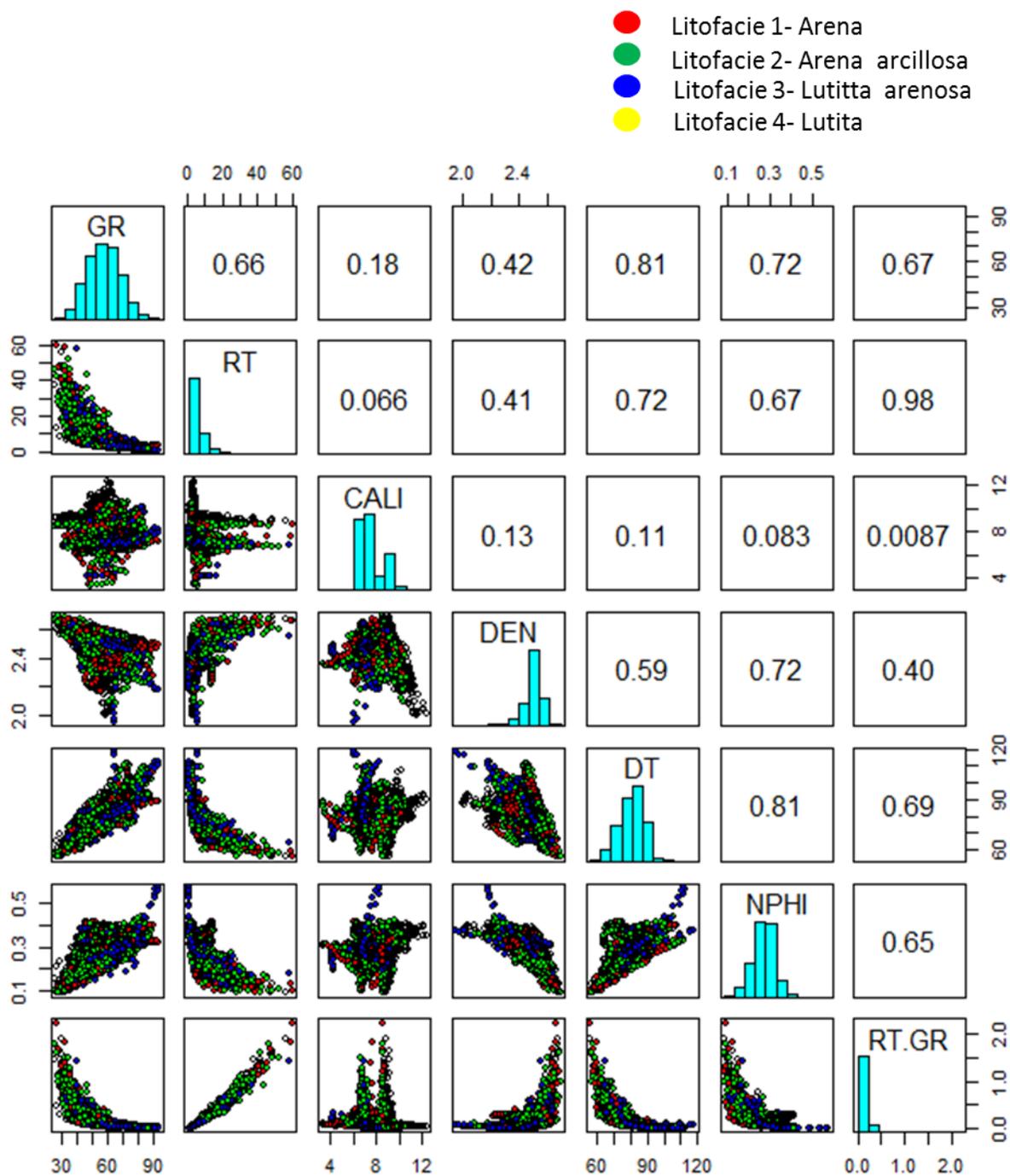


Figura B.10 Gráfico de dispersión para todos los pozos.

III. Análisis por pozo

a) Gráfico de caja (boxplot)

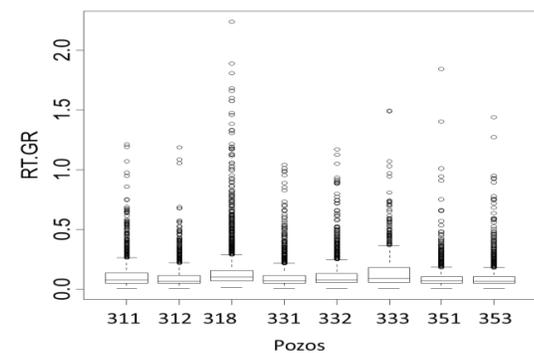
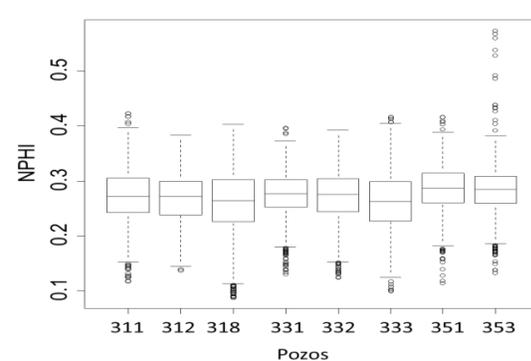
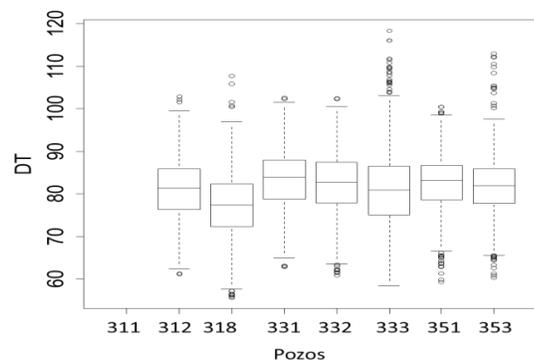
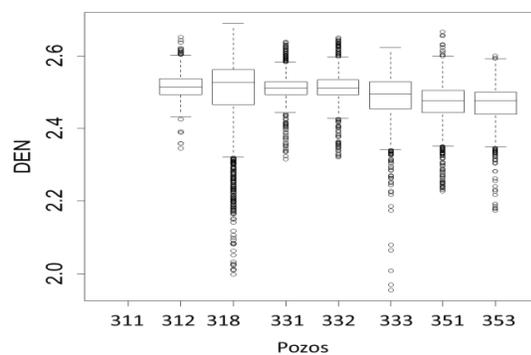
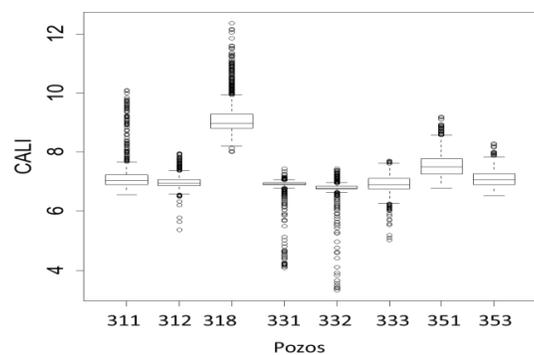
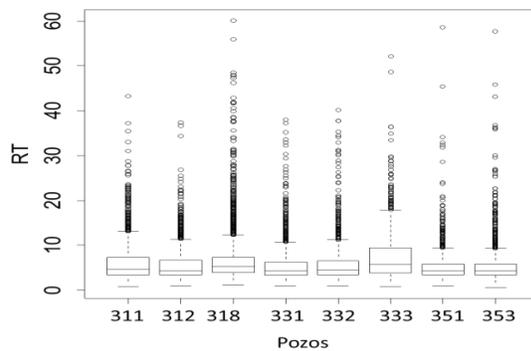
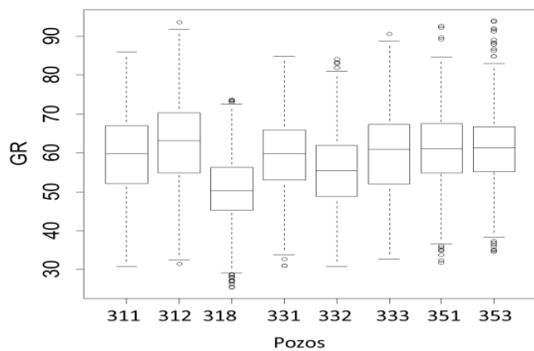


Figura B.11. Gráfico de caja por pozo

b) Gráficos de dispersión por pozo.

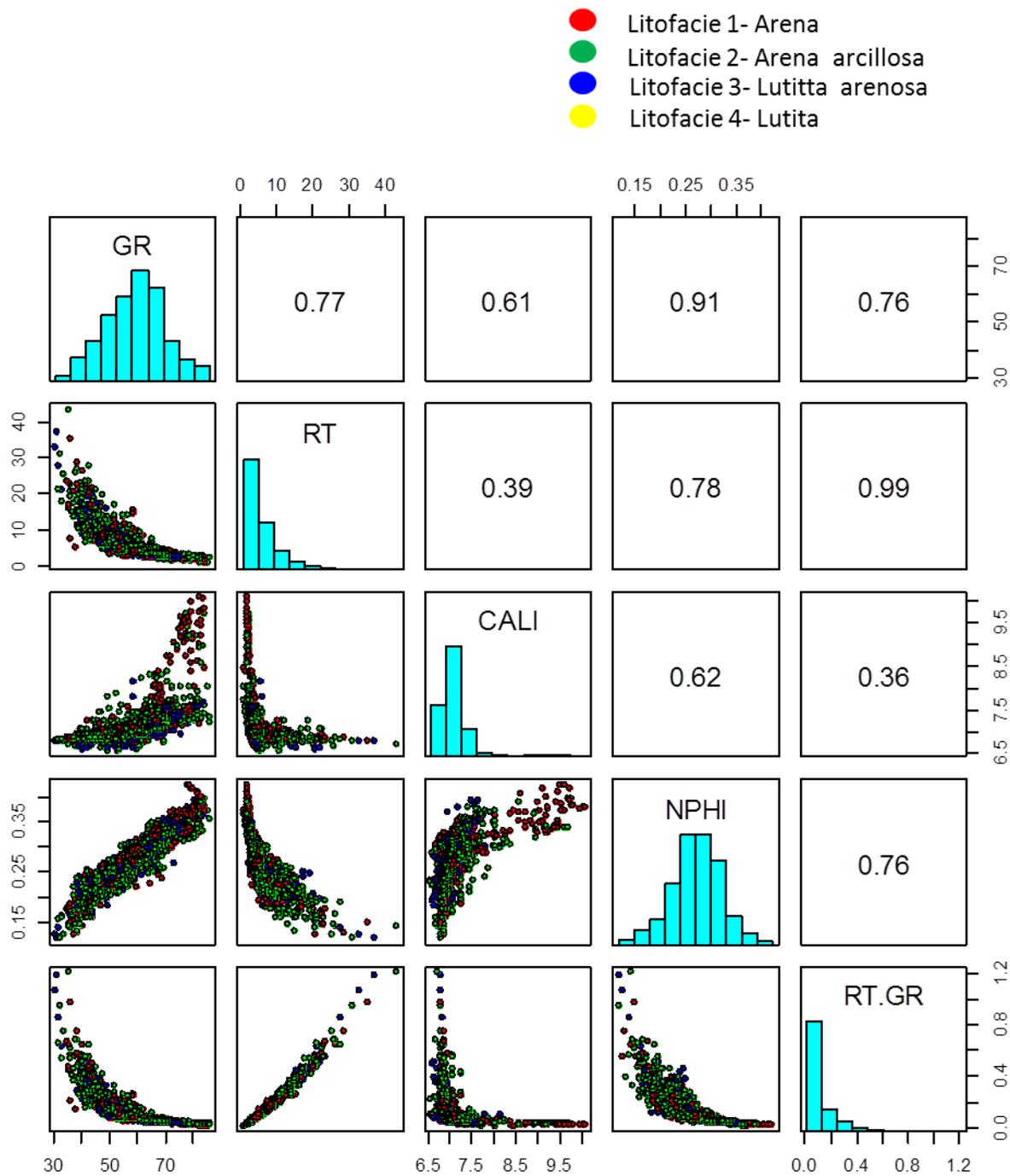


Figura B.12 Gráfico de dispersión - Pozo 311

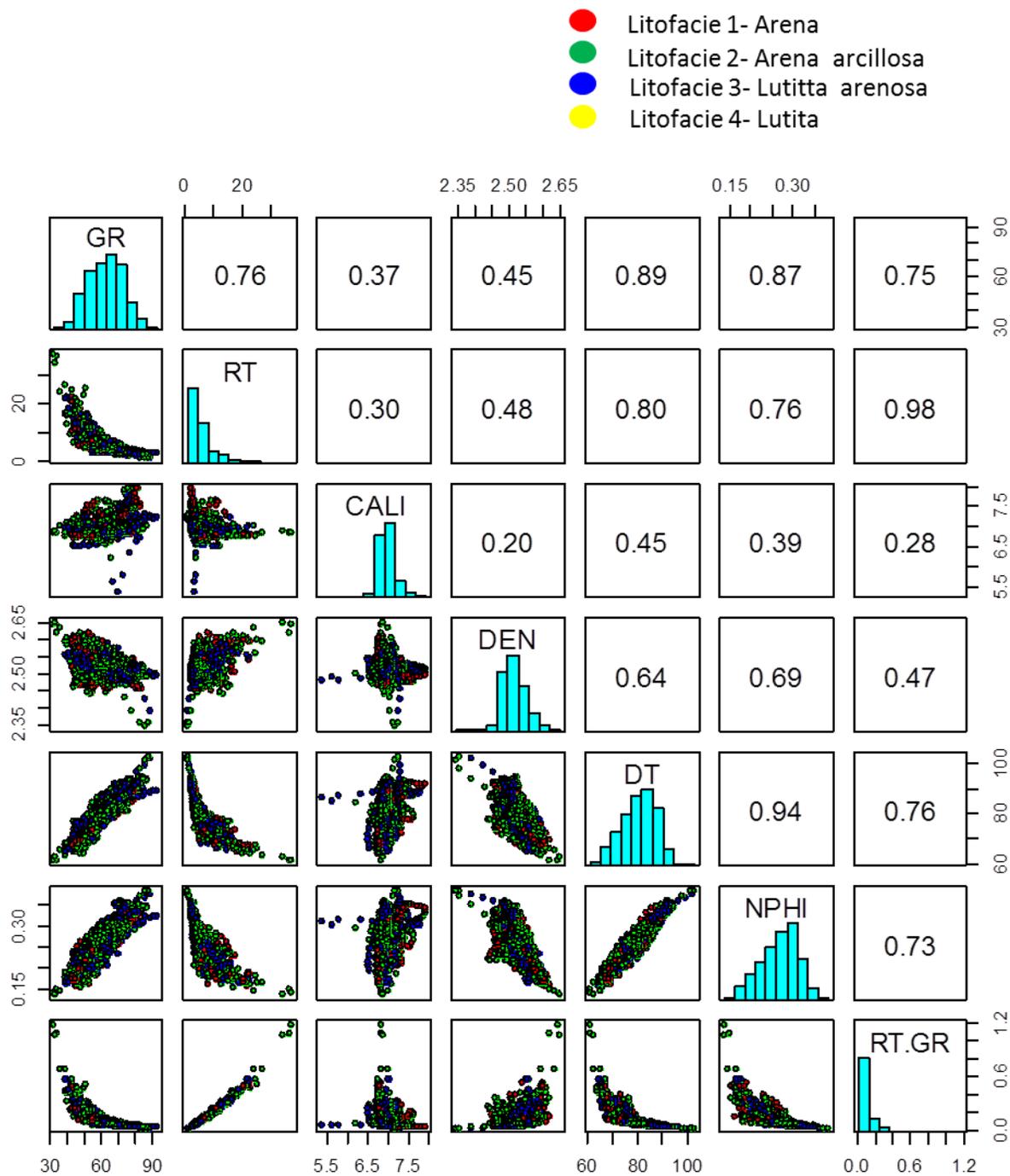


Figura B.13. Gráfico de dispersión - Pozo 312

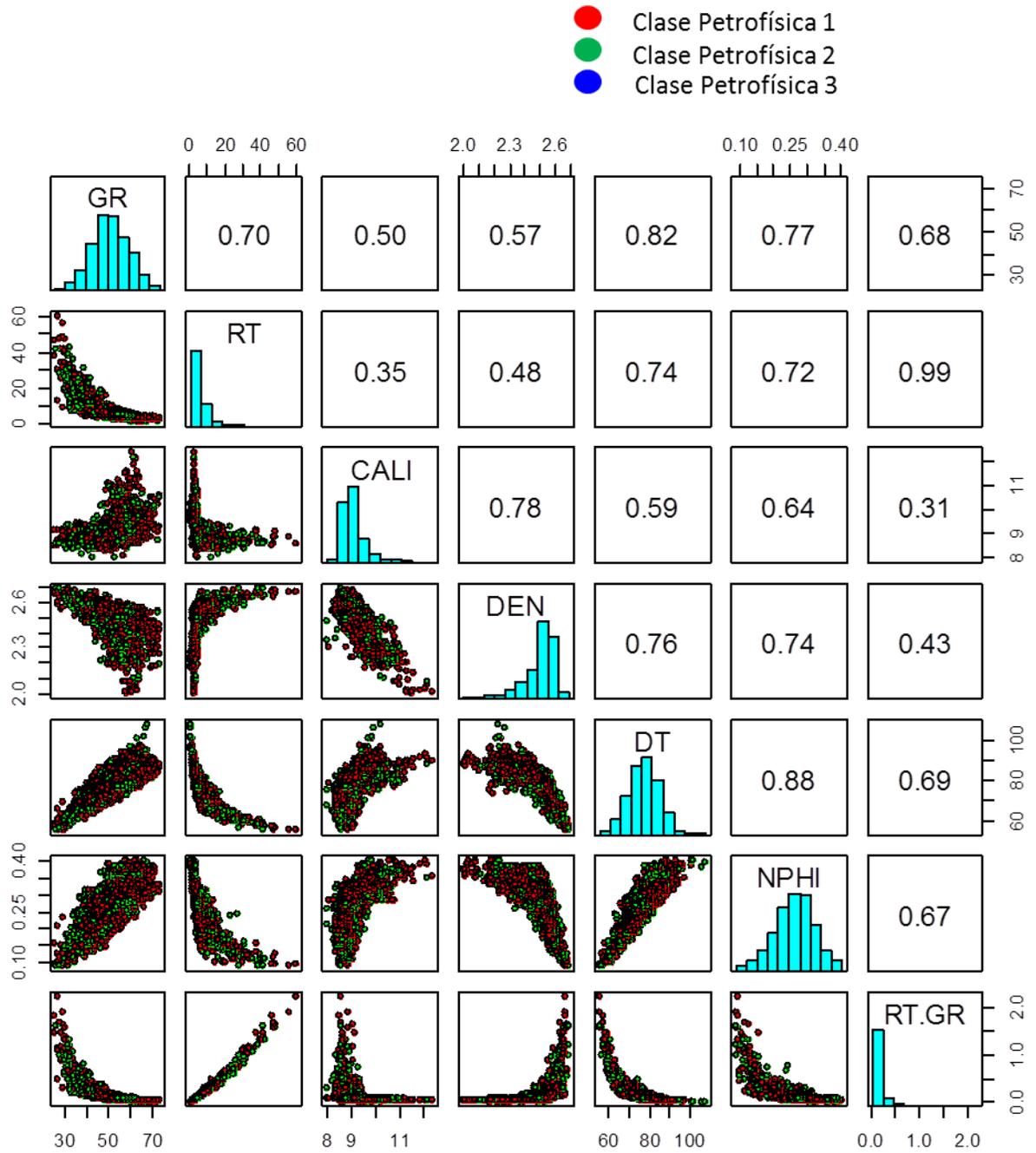


Figura B.14. Gráfico de dispersión - Pozo 318

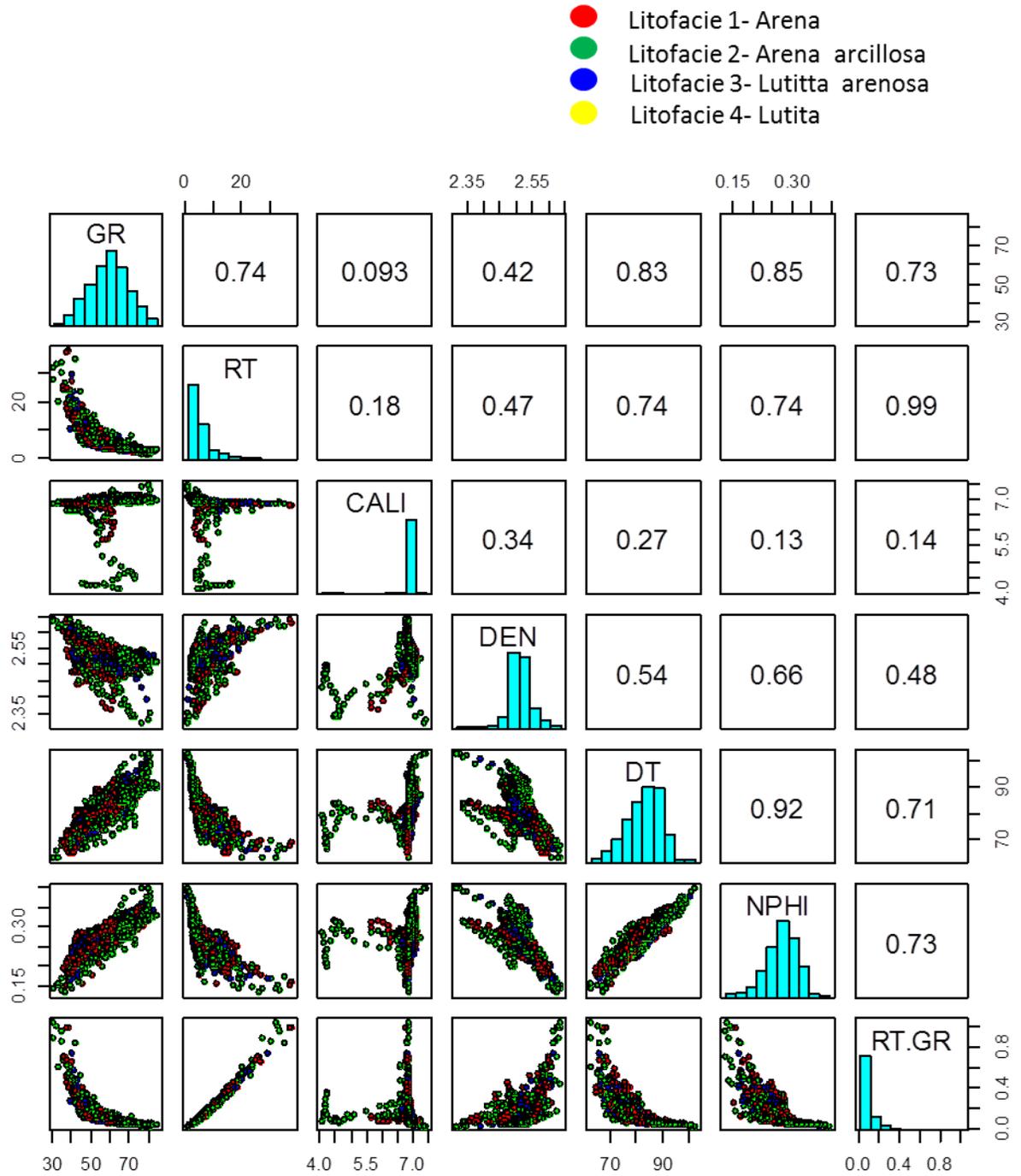


Figura B.15. Gráfico de dispersión - Pozo 331

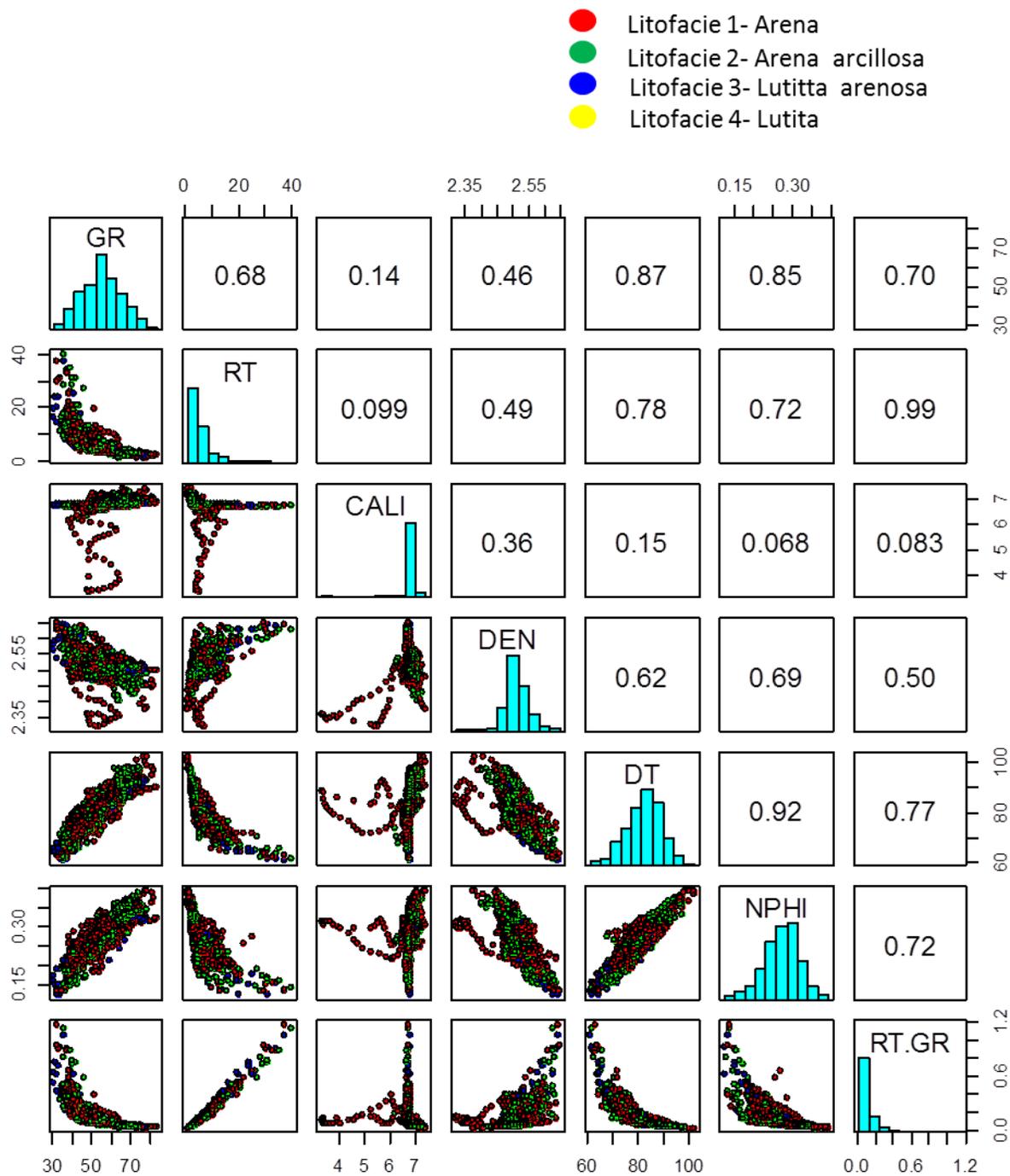


Figura B.16. Gráfico de dispersión - Pozo 332

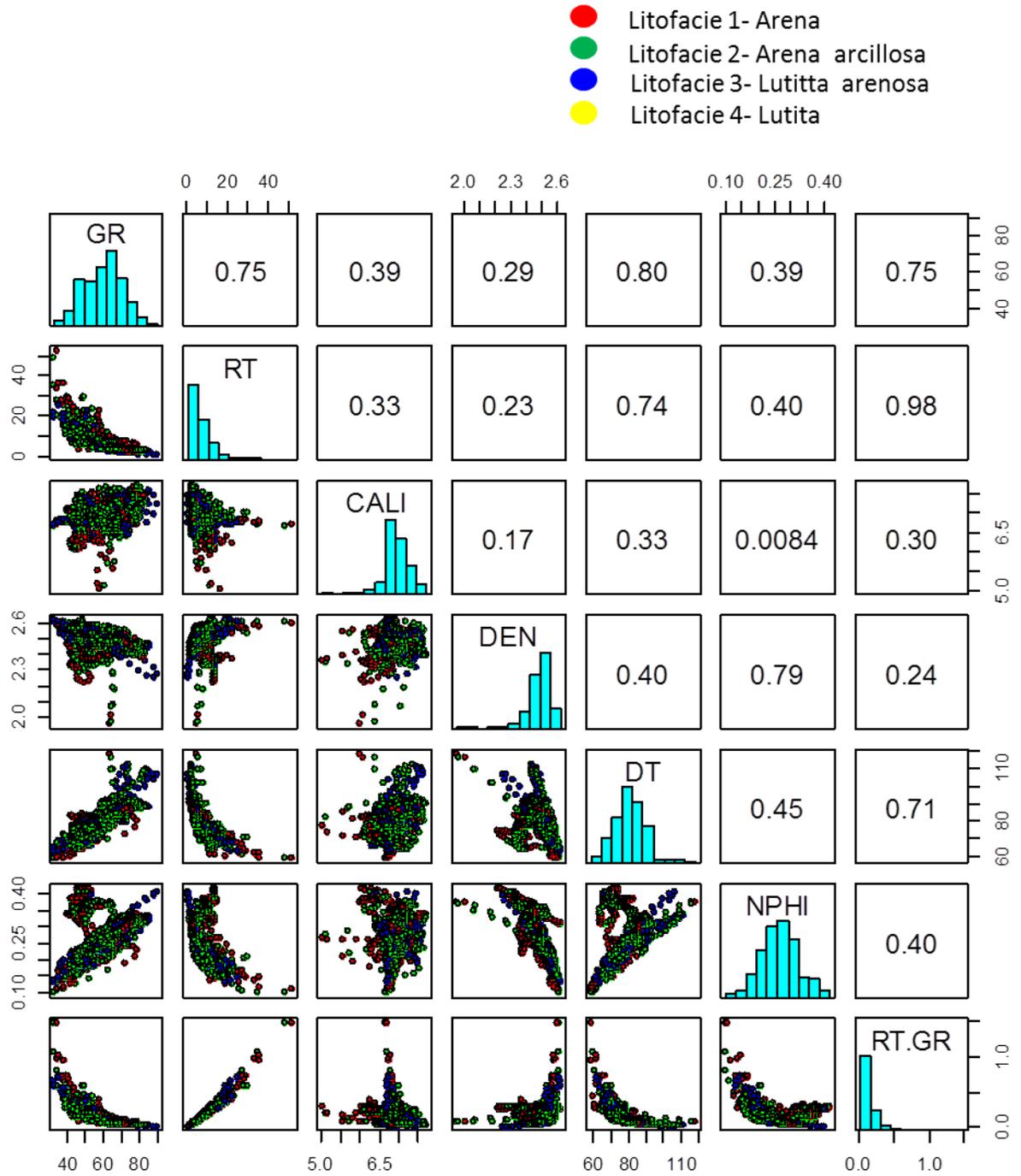


Figura B.17. Gráfico de dispersión - Pozo 333

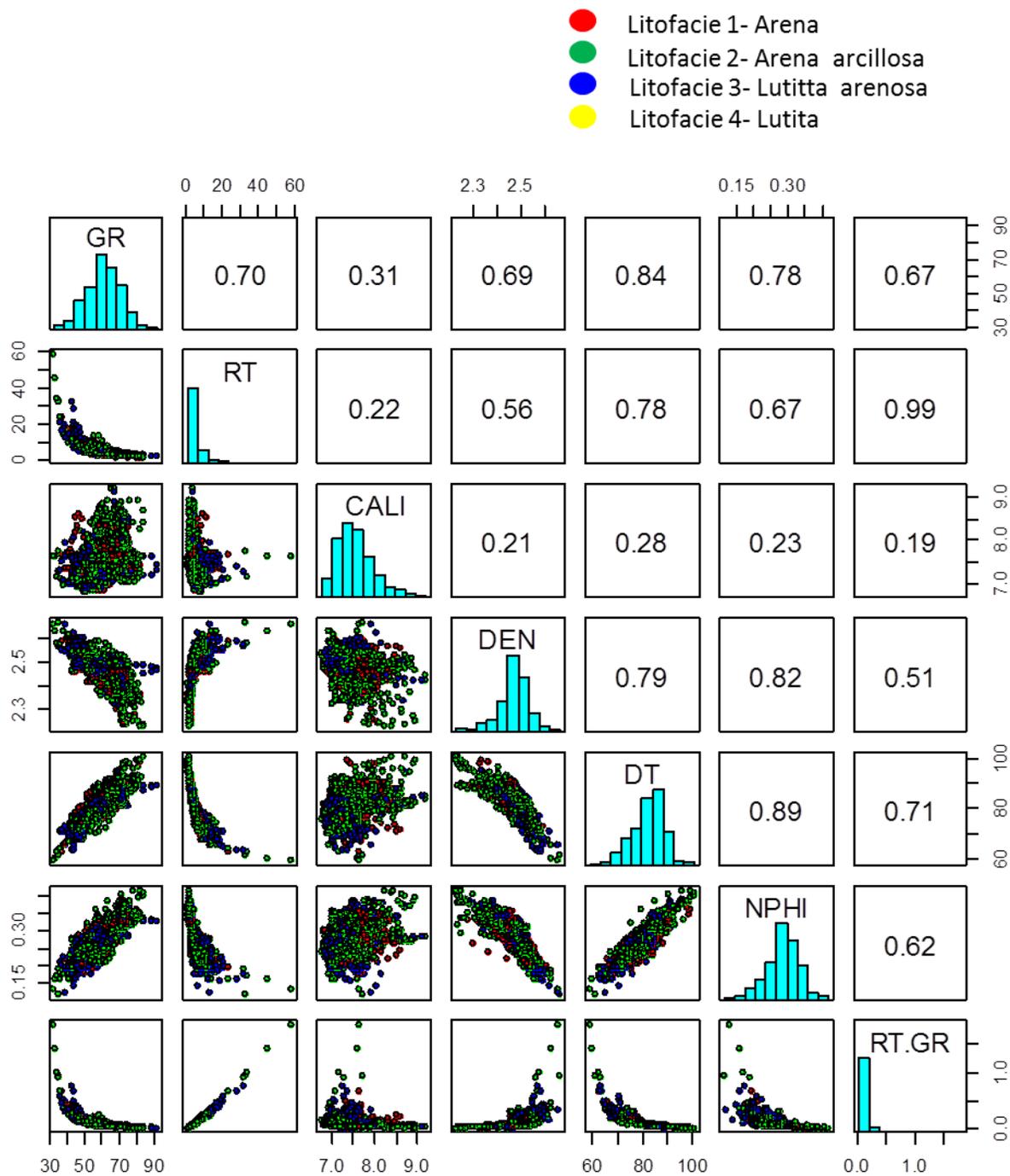


Figura B.18. Gráfico de dispersión - Pozo 351

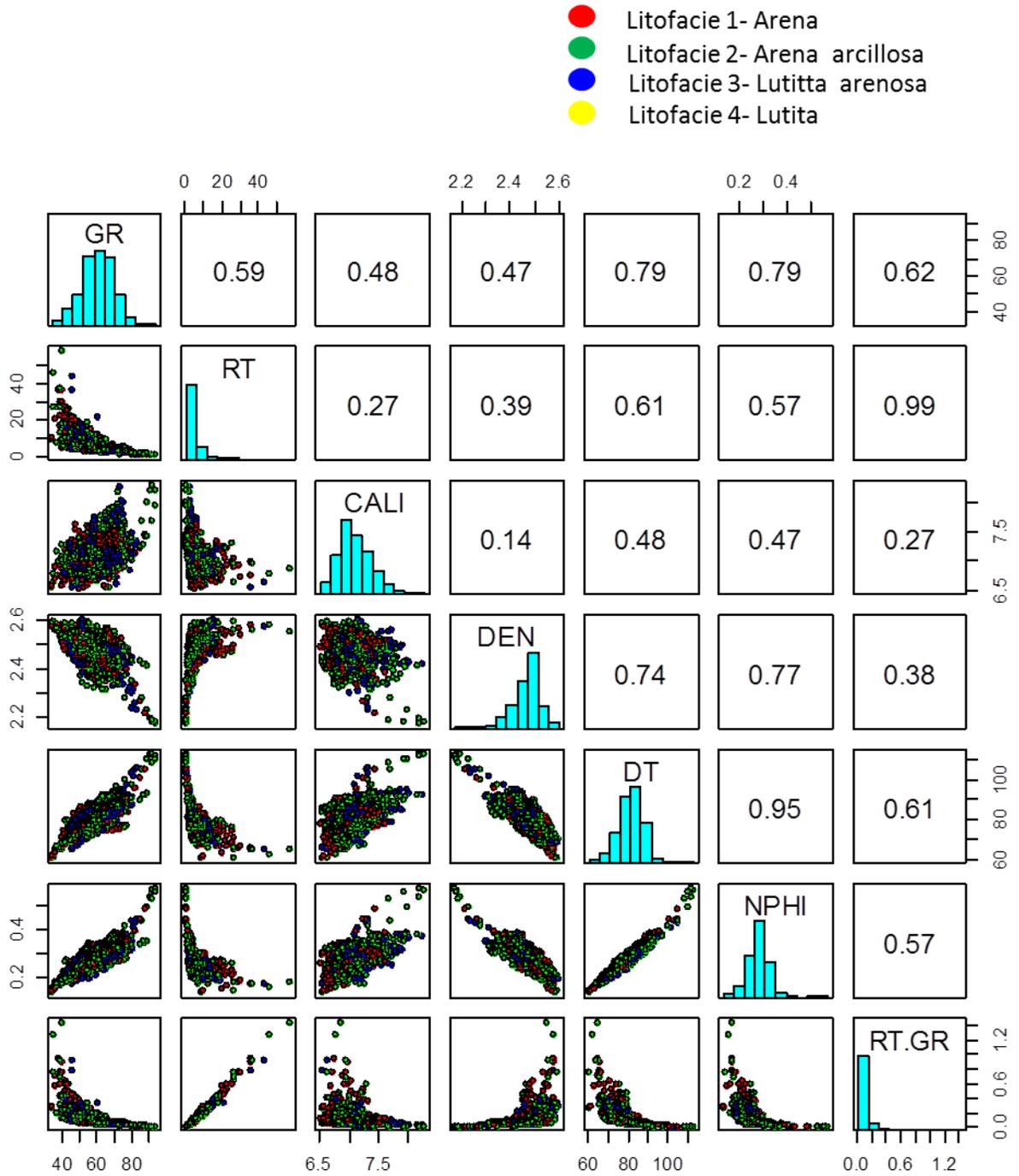


Figura B.19. Gráfico de dispersión - Pozo 353

REFERENCIAS

1. Ackley, D.H., Hinton, G.E. y Sejnowski, T.J. 1985. A Learning Algorithm for Boltzmann Machines. *Cognitive Science* 9 (1): 147-169.
2. Aguilar Jaramillo, J. 2007. Apuntes de Inteligencia Artificial, Laboratorio: Redes con Conexión Hacia Adelante. Pontificia Universidad Católica del Ecuador, Facultad de Ingeniería. http://www.flacsoandes.org/comunicacion/aaa/imagenes/publicaciones/pub_30.pdf (descargado 01 de enero de 2013).
3. Aguilar Sánchez, A.R. 2008. Las Redes Neuronales Artificiales como una Herramienta Alternativa o Complementaria a los Métodos de Clasificación Tradicionales. Tesina de Licenciatura, Universidad Nacional Autónoma de México, Facultad de Estudios Superiores-Acatlán.
4. Alpana, B. 2002. Reservoir Properties from Well Logs Using Neural Networks. Doktor Ingeniør thesis, Norwegian University of Science and Technology. http://www.ipt.ntnu.no/~bjornu/Alpana_Bhatt.pdf (descargado 01 de enero de 2013).
5. Anand, R., Mehrotra, K., Mohan, C.K. et al. 1995. Efficient Classification for Multiclass Problem Using Modular Neural Networks. *IEEE Transactions on Neural Networks* 6 (1): 117-124. <http://dx.doi.org/10.1109/72.363444>.
6. Anderson, D. y McNeill, G. 1992. Artificial Neural Networks Technology: A DACS State-of the-Art Report. ELIN: A011, Data & Analysis Center for Software, Rome Laboratory, Griffiss AFB, New York (20 de agosto 1992). http://andrei.clubcisco.ro/cursuri/f/f-sym/5master/aac-nnga/AI_neural_nets.pdf (descargado 01 de enero de 2013).
7. Arm Geophysics, Surface & Borehole Services. 2013. Typical Well Responses. <http://www.armgeophysics.net/Documents/Well%20Log%20Response%20Chart.pdf> (descargado 01 de enero de 2013).
8. Artiga Landaverde, A.J. y Escobar Flores, G.F. 2006. Sistema de Reconocimiento de Patrones Rítmicos para Batería Acústica. Tesis de Licenciatura, Universidad Don Bosco, Facultad de Ingeniería, Ciudadela Don Bosco, El Salvador. http://rd.udb.edu.sv:8080/jspui/bitstream/123456789/295/1/038310_tesis.pdf (descargado 01 de enero de 2013).
9. Bernal García, J.J., Martínez María-Dolores, S.M. y Sánchez García, J.F. 2004. Modelización de los Factores más Importantes que Caracterizan un Sitio en la Red. XVII Jornadas de la Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa, Universidad de Murcia, septiembre 16-17. http://www.uv.es/asepuma/XII/comunica/bernal_martinez_sanchez.pdf (descargado 01 de enero de 2013).
10. Bishop, C.M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.
11. Block, H.D. 1962. The Perceptron: A Model for Brain Functioning I. *Reviews of Modern Physics* 34 (1): 123-135. <http://dx.doi.org/10.1103/RevModPhys.34.123>.

12. Breiman, L. 1996. Bagging Predictors. *Machine Learning* 24 (2): 123-140. <http://dx.doi.org/10.1023/A:1018054314350>.
13. Caers, J. 2005. *Petroleum Geostatistics*. Richardson, Texas: SPE.
14. Carpenter, G.A. y Grossberg, S. 1986. Absolutely Stable Learning of Recognition Codes by A Self-Organizing Neural Network. En *Neural Networks for Computing, AIP Conf. Proc.*, ed. J.S. Denker, Vol. 151, 77-85. <http://dx.doi.org/10.1063/1.36223>.
15. Casar González, R., Díaz Viera, M.A., Hernández Maldonado, V. et al. 2012. Metodología para obtener Modelos Geológicos-Petrofísicos Basados en la Aplicación de Métodos Geoestadísticos.
16. Casar González R., Díaz Viera M.A., Méndez Venegas J., Hernández Maldonado V.M., Checa Rojas P., Mendoza Torres F., Torres Vargas R., Cruz Castillo M., Acosta Ángeles A. 2013. Modelo Geológico-Petrofísico del Intervalo de la Prueba Piloto de Inyección de CO₂ en la Macropera 331, Campo Coyotes, Cuenca de Chicontepec. Registro en trámite, Instituto Nacional del Derecho de Autor, Agosto 2013.
17. Castillo Villegas, A. 2007. La Aplicación de Redes Neuronales en la Distribución de Planta. Tesis de Licenciatura, Universidad Autónoma del Estado de Hidalgo, Instituto de Ciencias Básicas e Ingeniería, Pachuca de Soto, Hidalgo.
18. Cho, S., Jang, M. y Chang, S. 1997. Virtual Sample Generation Using a Population of Networks. *Neural Processing Letters* 5 (2): 21-27. <http://dx.doi.org/10.1023/A:1009653706403>.
19. Correa Martínez, J. y Porras Espinosa, J. 1999. Redes Neuronales: Estudio y Aplicación al Reconocimiento de Caracteres y Control. Tesis de Licenciatura, Universidad Autónoma Metropolitana, Unidad Iztapalapa, Ciencias Básicas e Ingeniería Electrónica.
- Cosentino, L. 2001. *Integrated Reservoir Studies*. Paris: Editions Technip.
20. Crain, E.R. 2013. Welcome to Crain's Petrophysical Handbook. Online Shareware Petrophysics Training and Reference Manual: Chapter and Topics. <http://spec2000.net/01-index.htm> (Fecha de acceso 10 de enero de 2013).
21. Corchado, J., Díaz, F., Borrajo, L., Fernández, F. 2000. *Redes Neuronales Artificiales: un enfoque práctico*. Servicio Publicaciones de la Universidad de Vigo.
22. Cuevas L. 2005. *Avances del Proyecto Lankahuasa*, Colegio de Ingenieros Petroleros de México, 2005.
23. Curso-R. 2009. <http://curso-r-uah2009.wikispaces.com> (Fecha de acceso 10 de enero de 2013).
24. Deutsch, C.V. 2002. *Geostatistical Reservoir Modeling*. Oxford, UK: Oxford University Press.
25. Díaz Viera, M.A. 2002. *Notas del Curso: Geoestadística Aplicada*. Universidad Nacional Autónoma de México, Instituto de Geofísica e Instituto de Geofísica y Astronomía de Cuba.
26. Díaz Viera, M.A. y Casar González, R. 2004. *Presentaciones del Curso: Geoestadística Aplicada*. México, D.F.: Instituto Mexicano del Petróleo 2004.
27. Díaz Viera, M.A. y Casar González, R. 2006 *Manual del Instructor del Curso: Modelación Geológica-Petrofísica de Yacimientos*. México, D.F.: Instituto Mexicano del Petróleo.
28. Dresser Industries, Inc. 1979. *Log Interpretation Charts*. Houston: Dresser Industries, Inc.

-
29. Duda, R.O., Hart, P.E. y Stork, D. 2001 *Pattern Classification*, second edition. New York: Probabilistic and Statistic, John Wiley and Sons.
30. Dunne, R.A. 2007. *A Statistical Approach to Neural Networks for Pattern Recognition*. Chichester, UK: Wiley Series in Computational Statistics, Wiley-Interscience.
31. Enikanselu, P.A. y Ojo, A.O. 2012. Statistical Analysis and Evaluation of Lithofacies from Wireline Logs Over 'Beleema' Field, Niger Delta, Nigeria. *JPGE* 3 (2): 26–34. <http://www.academicjournals.org/jpge/PDF/pdf2012/1%20Feb/Enikaselun%20and%20Ojo.pdf> (descargado 01 de enero de 2013).
32. Fang, J.H., Chen, H.C., Shultz, A.W. et al. 1992. Computer-Aided Well Log Correlation (1). *AAPG Bull.* 76 (3): 307-317.
33. Faria, C., Agnelo, J. y Bettini, C. 2005. The use of well logs in logfacies modeling – example in the Namorado Field, Campos Basin, Brazil, 2005, Ninth International Congress of the Brazilian Geophysical Society, Brasil. <http://www.earthdoc.org/publication/publicationdetails/?publication=40429> (descargado 15 de febrero de 2013)
34. Fausett, L.V. 1994. *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*. Englewood Cliffs, New Jersey: Prentice-Hall.
35. Ferraz, N.I. y Garcia, A.C.B. 2005. Lithofacies Recognition Hybrid Bench. En Proc., HIS'05, 5th International Conference on Hybrid Intelligent Systems, Rio de Janeiro, noviembre 6-9, Vol. 1 87 – 91. <http://dx.doi.org/10.1109/ICHIS.2005.70>
36. Flórez López, R. y Fernández Fernández, J.M. 2008. *Las Redes Neuronales Artificiales: Fundamentos teóricos y aplicaciones Prácticas*. Oleiros, La Coruña, España: Netbiblo.
37. Freund, Y. y Schapire, S. 1996. Experiments with a New Boosting Algorithm. En: *Machine Learning: Proceedings of the Thirteenth International Conference (ICML '96)*, Julio 3-6, Bari, Italia, ed. L. Saitta, 149-156, San Francisco, California: Morgan Kaufmann.
38. Fukushima, K. 1975. Cognitron: A Self-Organizing Multilayered Neural Network. *Biological Cybernetics* 20 (3-4): 121-136. <http://dx.doi.org/10.1007/BF00342633>.
39. Fyfe, C. 1996. *Artificial Neural Networks*, edition 1.1. http://cdn.preterhuman.net/texts/science_and_technology/artificial_intelligence/Artificial%20Neural%20Networks%20-%20Colin%20Fyfe.pdf (descargado 01 de enero de 2013).
40. Galloway, W.E. 1998. Siliciclastic Slope and Base-of-Slope Depositional Systems: Component Facies, Stratigraphic Architecture, and Classification. *AAPG Bull.* 82 (4): 569-595.
41. García González, Y. 2011. *Modelos y Algoritmos para Redes Neuronales Recurrentes Basadas en Wavelets Aplicados a la Detección de Intrusos*. Tesis de Maestría, Universidad de las Américas Puebla, Cholula, Puebla, México.

-
42. Geman, S., Bienenstock, E. y Doursat, R. 1992. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* 4 (1): 1-58.
<http://stuff.mit.edu/afs/athena.mit.edu/course/6/6.435/OldFiles/www/Geman92.pdf> (descargado 01 de enero de 2013).
43. Gill, D., Shamrony, A. y Fligelman, H. 1993. Numerical Zonation of Log Suites and Logfacies Recognition by Multivariate Clustering. *AAPG Bull.* 77 (10): 1781-1791.
44. González G., N.E. y Méndez R., J.J. 2007. Clasificación de Facies Sedimentarias en Yacimientos Geológicamente Complejos Aplicando Redes Neuronales Probabilísticas (PNN). En IX Congreso Geológico Venezolano, octubre 25-27, Universidad Central de Venezuela, Caracas Venezuela.
45. Grossberg, S. 1972a. A Neural Theory of Punishment and Avoidance, I: Qualitative Theory. *Mathematical Biosciences* 15 (1-2): 39-67. [http://dx.doi.org/10.1016/0025-5564\(72\)90062-4](http://dx.doi.org/10.1016/0025-5564(72)90062-4).
46. Grossberg, S. 1972b. A Neural Theory of Punishment and Avoidance, II: Quantitative Theory. *Mathematical Biosciences* 15 (3-4): 253-285. [http://dx.doi.org/10.1016/0025-5564\(72\)90038-7](http://dx.doi.org/10.1016/0025-5564(72)90038-7).
47. Grossberg, S. 1976. Adaptive Pattern Classification and Universal Recoding: I. Parallel Development and Coding of Neural Feature Detectors. *Biological Cybernetics* 23 (3): 121-134. <http://dx.doi.org/10.1007/BF00344744>.
48. Gutiérrez Martín, F. 2000. *Redes Neuronales Artificiales*. <http://thales.cica.es/rd/Recursos/rd98/TecInfo/07/> (Fecha de acceso 10 de enero de 2013).
49. Hagan, M.T., Demuth, H.B. y Beale, M. 1996. *Neural Network Design*. Boston, Massachusetts: PWS Publishing Co.
50. Hagan, M.T., Demuth, H.B. y Beale, M. 2009. *Neural Network Toolbox 6 User's Guide*. Natick, Massachusetts: The MathWorks, Inc.
51. Hashem, S. 1997. Optimal Linear Combination of Neural Networks. *Neural Networks* 10 (4): 599-614. [http://dx.doi.org/10.1016/S0893-6080\(96\)00098-6](http://dx.doi.org/10.1016/S0893-6080(96)00098-6).
52. Haykin, S.S. 1999. *Neural Networks: A Comprehensive Foundation*, second edition. Upper Saddle River, New Jersey: Prentice Hall.
53. Hilera González, J. y Martínez Hernández, V. 1995. *Redes Neuronales Artificiales: Fundamentos, Modelos y Aplicaciones*. Madrid: RA-MA.
54. Hopfield, J. 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. USA* 79 (8): 2554-2558.
<http://www.pnas.org/content/79/8/2554.full.pdf> (descargado 01 de enero de 2013).
55. Hornik, K.M., Stinchcombe, M. y White, H. 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* 2 (5): 359-366. http://weber.ucsd.edu/~hwhite/pub_files/hwcv-028.pdf (descargado 01 de enero de 2013).
56. Hristev, R.M. 1998. *The ANN Book*, edition1.

http://www.gsd.ece.buap.mx/DocumentosSistemasDigitales/NeuralNetworks/Hritsev_The_ANN_Book.pdf
(descargado 01 de enero de 2013).

57. Jiménez Meneses, G. y Sandoval Chávez, D.A. 2009. Estrategias de Explotación de Campos Petroleros con Características Similares a Chicontepec. Tesis de Licenciatura, Universidad Nacional Autónoma de México, Facultad de Ingeniería, México, D.F.
58. Jolliffe, I.T. 2002. Principal Component Analysis, second edition. New York: Springer Series in Statistics, Springer-Verlag.
59. Jones, I.F. 2010. An Introduction to Velocity Model Building. Houten, Netherlands: European Association of Geoscientists & Engineers Publications.
60. Jordan, M.I. y Jacobs, R.A. 1994. Hierarchical Mixtures of Experts and the EM Algorithm., *Neural Computation* 6 (2): 181-214. <http://dx.doi.org/10.1162/neco.1994.6.2.181>.
61. Kala, R. 2013. Rahul Kala, School of Cybernetics, University of Reading, UK [página personal]. <http://rkala.99k.org/> (Fecha de acceso 10 de enero de 2013).
62. Karypis, G., Han, E.H. y Kumar, V. 1999. Chamedeuleon: Hierarchical Clustering Using Dynamic Modeling. *Computer* 32 (8): 68-75. <http://dx.doi.org/10.1109/2.781637>.
63. Kelkar, M., Perez, G. y Chopra, A. 2002. Applied Geostatistics for Reservoir Characterization. Richardson, Texas: SPE.
64. Kohonen, T. 1982a. Analysis of a Simple Self-Organizing Process. *Biological Cybernetics* 44 (2): 135-140. <http://dx.doi.org/10.1007/BF00317973>.
65. Kohonen, T. 1982b. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* 43 (1): 59-69. <http://dx.doi.org/10.1007/BF00337288>.
66. Kohonen, T. 1989. Self-Organization and Associative Memory, third edition. Berlin: Springer Series in Information Sciences, Springer-Verlag.
67. Kohonen, T. 1995. Self-Organizing Maps. Berlin: Springer Series in Information Sciences, Springer-Verlag.
68. Kohonen, T., Hynninen, J., Kangas, J. et al. 1995. SOM_PAK: The Self-Organizing Map Program Package Version 3.1 (April 7, 1995). SOM Programming Team of the Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finlandia. http://www.isegi.unl.pt/ensino/docentes/fbacao/som_pak.pdf (descargado 01 de enero de 2013).
69. Kumar, B. y Kishore, M. 2006. Electrofacies Classification – A Critical Approach. En 6th International Conference & Exposition on Petroleum Geophysics, enero 9-11, Kolkata India, 822-825. http://www.spgindia.org/conference/6thconf_kolkata06/360.pdf (descargado 01 de enero de 2013).
70. León González, A., Llinás Solano, H. y Tilano, J. 2008. Análisis Multivariado Aplicando Componentes Principales al Caso de los Desplazados. *Ingeniería y Desarrollo* 23: 119-142. <http://www.redalyc.org/articulo.oa?id=85202310> (descargado 01 de enero de 2013).

71. López Gómez, X. 2001. Aplicación de la Estadística en la Toma de Decisiones para el Reconocimiento de Patrones Mediante Redes Neuronales. Tesis de Licenciatura, Universidad Nacional Autónoma de México, Facultad de Ciencias, México, D.F.
72. Ma, Y.Z. 2011. Lithofacies Clustering Using Principal Component Analysis and Neural Network: Applications to Wireline Logs. *Mathematical Geosciences* 43 (4): 401-419. <http://dx.doi.org/10.1007/s11004-011-9335-8>.
73. Marín Diazaraque, J.M. 2012. Introducción a las Redes Neuronales Aplicadas: Curso de EXPERTOS de U.C.M. (2012). Universidad Carlos III de Madrid.
<http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Expertos/PresNN12.pdf> (descargado 01 de enero de 2013).
74. Martín del Brio, B. y Sanz Molina, A. 1997. *Redes Neuronales y Sistemas Borrosos: Introducción, Teoría y Práctica*. Madrid: Ra-Ma.
75. Martínez Luaces, M. 2006-2007. *Redes Neuronales Artificiales Modulares: Una Posible Especificación y su Aplicación a Problemas Tecnológicos así como al Modelado e Implementación de Funciones Cognitivas*. Memoria del trabajo tutelado, Universidad Politécnica de Madrid, Facultad de Informática, Madrid, España.
http://www.dlsiis.fi.upm.es/docto_lsiis/Trabajos20062007/Milton.pdf (descargado 01 de enero de 2013).
76. Martínez, J. 2012. Proyecto CEASE. Universidad de Valencia. España.
<http://www.uv.es/ceaces/multivari/cluster.html> (Fecha de acceso octubre de 2012)
77. McAuley, D. y Simon, D. 1999. *The Back Propagation Network: Learning by Example*.
<http://itee.uq.edu.au/~cogs2010/cmc/chapters/BackProp/> (Fecha de acceso 10 de enero de 2013).
78. Mehotra, K., Mohan, C. K. y Ranka, S. 1997. *Elements of Artificial Neural Networks*. Cambridge, Massachusetts: The MIT Press.
79. Mejía Sánchez, J.A. 2004. *Sistema de Detección de Intrusos en Redes de Comunicaciones Utilizando Redes Neuronales*. Tesis de Licenciatura, Universidad de las Américas Puebla, Escuela de Ingeniería, Cholula, Puebla, México.
80. Mirowski, P. y McCormick, D. 2003. Inferring Lithofacies from Well Logs by Applying Hybrid Neural Network-Hidden Markov Model Classifiers. En *AAPG International Conference*, septiembre 21-24. Barcelona, España.
http://www.searchanddiscovery.com/abstracts/pdf/2003/intl/short/ndx_82901.pdf (descargado 01 de enero de 2013).
81. Montaña Moreno, J.J. 2002. *Redes Neuronales Artificiales Aplicadas al Análisis de Datos*. Tesis Doctoral, Universitat de les Illes Balears, Facultad de Psicología, Palma de Mallorca, España.
<http://www.tdx.cat/bitstream/handle/10803/9441/tjimm1de1.pdf;jsessionid=4857633FD9F214DB2E87544FA4301930.tdx2?sequence=1> (Fecha de acceso 10 de enero de 2013).
82. Moore, R.C. 1949. Meaning of Facies. En *Sedimentary Facies in Geologic History Geology: Conference at Meeting of The Geological Society of America*, noviembre 11, New York, New York, C.R. Longwell, chairman, 1-34. *GSA Memories* 39. <http://dx.doi.org/10.1130/MEM39-p1>.

- 83.** Moré, J.J. 1978. The Levenberg-Marquardt Algorithm: Implementation and Theory. En Numerical Analysis: Proc. of the Biennial Conference, junio 28-julio 1 1977, Dundee, UK, ed. G.A. Watson, Vol. 630, 105-116. Berlin: Lecture Notes in Mathematics, Springer-Verlag. <http://dx.doi.org/10.1007/BFb0067700>.
- 84.** Naftaly, U., Intrator, N. y Horn, D. 1997. Optimal Ensemble Averaging of Neural Networks. *Network: Comput. Neural Syst.* 8: 283-296.
<http://neuron.tau.ac.il/~horn/publications/optimal.pdf> (descargado 01 de enero de 2013).
- 85.** Narendra, K.S. y Thathachar, M. 1974. Learning Automata - A Survey. *IEEE Transactions on Systems, Man and Cybernetics SMC-4* (4): 323-334.
<http://dx.doi.org/10.1109/TSMC.1974.5408453>.
- 86.** Narendra, K.S. y Parthasarathy, K. 1990. Identification and Control of Dynamical Systems Using Neural Networks. *IEEE Transaction Neural Networks* 1 (1): 4-27.
<http://dx.doi.org/10.1109/72.80202>.
- 87.** Nguyen, D. y Widrow, B. 1990. Improving the Learning Speed of 2-Layer Neural Networks by Choosing Initial Values of the Adaptive Weights. Proc., IEEE International Joint Conference on Neural Networks, Washington D.C. Vol. 3, 21-26. <http://dx.doi.org/10.1109/IJCNN.1990.137819>.
- 88.** Ochoa, S. 2003. Resolución de Redes de Distribución Hidráulica Mediante Redes Neuronales Artificiales. Antioquia, Colombia: Escuela de Ingeniería de Antioquia, Ingeniería Civil. <http://fluidos.eia.edu.co/hidraulica/articulos/flujoentuberias/neuronal/neuronal.htm> (Fecha de acceso 10 de enero de 2013).
- 89.** Peña, D. 2002. Análisis de Datos Multivariantes. Madrid: McGraw-Hill/Interamericana de España.
- 90.** Popielski, A.C., Heidari, Z. y Torres-Verdín C. 2012. Rock Classification from Conventional Well Logs in Hydrocarbon-Bearing Shale. Artículo SPE 159255, presentado en SPE Annual Technical Conference and Exhibition, San Antonio, Texas, octubre 8-10. <http://dx.doi.org/10.2118/159255-MS>.
- 91.** Poulton, M.M. 2001. Computational Neural Networks for Geophysical Data Processing. New York: Handbook of Geophysical Exploration. Seismic Exploration, Pergamon.
- 92.** Robert, J. y Birkenhead, R. ed. 2001. Developments in Soft Computing. Berlin: Advances in Soft Computing, Springer-Verlag.
- 93.** Rodríguez Rojas, O. 2009. Análisis de Componentes Principales. http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/cap2.143111051.pdf (descargado 01 de enero de 2013).
- 94.** Rogers, S.J., Fang, J.H., Karr, C.L. et al. 1992. Determination of Lithology from Well Logs Using a Neural Network. *AAPG Bull.* 76 (5): 731-739.
- 95.** Rosenblatt, F. 1962. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington DC: Spartan Books.

96. Rui, X. y Wunsch, D. II. 2005. Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks* 16 (3): 645-678. <http://dx.doi.org/10.1109/TNN.2005.845141>.
97. Ruiz Cebrian, G. 2010. A Computer Aided Sequence Stratigraphy Approach: From Raw Data to Correlation and Sea-Level Variations. <http://www.docstoc.com/docs/25263572/A-COMPUTER-AIDED-SEQUENCE-STRATIGRAPHY-APPROACH-FROM-RAW-DATA> (Fecha de acceso 10 de enero de 2013).
98. Figueiras, M., Gallardo, P. 2003. Análisis Exploratorio de Datos, En Línea. Fecha de acceso 10 de abril de 2013). <http://www.5campus.com/leccion/aed>
99. Sánchez Romero, C.O. 2009 Influencia de Variables Macroclimáticas en la Sequía Meteorológica en el Sur del Perú. Tesina de Maestría, Universidad Politécnica de Valencia, Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos, Valencia, España. <http://riunet.upv.es/handle/10251/14545> (descargado 01 de enero de 2013).
100. Schapire, R.E. 1990. The Strength of Weak Learnability. *Machine Learning* 5 (2): 197-227. <http://dx.doi.org/10.1023/A:1022648800760>.
101. Schlumberger. 1975. A Guide to Well Site Interpretation for the Gulf Coast. Houston, Texas: Schlumberger Well Service. Inc.
102. Schlumberger. 2009. Petrel User Guide
103. Serra, O. y Abbott, H.T. 1980. The Contribution of Logging Data to Sedimentology and Stratigraphy. *SPE J.* 22 (1): 117-131. *SPE-9270-PA*. <http://dx.doi.org/10.2118/9270-PA>.
104. Sharkey, A.J.C. ed. 1999. Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems. New York: Perspectives in Neural Computing, Springer-Verlag.
105. Sharkey A.J.C., Sharkey N.E. y Chandroth, G.O. 1996. Diverse Neural Net Solutions to a Fault Diagnosis Problem. *Neural Computing & Applications* 4 (4): 218-227. <http://dx.doi.org/10.1007/BF01413820>.
106. Sharkey, A.J.C., Sharkey, N.E., Gerecke U. et al. 2000. The "Test and Select" Approach to Ensemble Combination. En: *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italia, junio 21–23, 2000 Proceedings*, ed. J. Kittler and F. Roli, Vol. 1857, 30-44. Berlin: Lecture Notes in Computer Science, Springer-Verlag. http://dx.doi.org/10.1007/3-540-45014-9_3.
107. Shatil, A.M.S. 2006. Bangla Optical Character Recognition Using Kohonen Network, 2006, Thesis BSc, BRAC University, Department of Computer Science and Engineering, Daca, Bangladés.
108. Stefanoic, P. y Kurasova, O. 2011. Visual Analysis of Self-Organizing Maps. *Nonlinear Analysis: Modelling and Control* 16 (4): 488-504. <http://www.lana.lt/journal/43/NA16409.pdf> (descargado 01 de enero de 2013).
109. Senenergy. 2005. Interactive Petrophysics V 3.1. Archivos de ayuda.
110. Tarrádez Gurrea, M. 2005. Análisis de Componentes Principales. http://www.uoc.edu/in3/emath/docs/Componentes_principales.pdf (descargado 01 de enero de 2013).
111. Universidad de Alcalá, Departamento de Teoría de la Señal y Comunicaciones, Grupo de Señales y Circuitos. 2012. Señales y Sistemas II: Zona de Apuntes. http://agamenon.tsc.uah.es/Asignaturas/it/ssii/zona_de_apuntes.htm (Fecha de acceso 10 de enero de 2013).

- 112.** Universidad de Granada, Facultad de Ciencias, Departamento de Estadística e Investigación Operativa. 2013. Contenidos de Fundamentos de Biología Aplicada I, Licenciatura en Biología. <http://www.ugr.es/~bioestad/TEMARIO%20FUND.htm> (Fecha de acceso 10 de enero de 2013).
- 113.** Vensanto, J., Himberg, J., Alhoniemi, E. et al. 1999. Self-Organizing Map in Matlab: the SOM Toolbox. Proc. of the Matlab DSP Conference 1999, Espoo, Finlandia, noviembre 16-17. <http://lib.tkk.fi/Diss/2002/isbn951226093X/article2.pdf> (descargado 01 de enero de 2013).
- 114.** Vensanto J., Himberg J., Alhoniemi E. et al. 2000. SOM Toolbox for Matlab 5. Report A57, SOM Toolbox Team, Hut, Finlandia.
- 115.** <http://www.cis.hut.fi/somtoolbox/package/papers/techrep.pdf> (descargado 01 de enero de 2013).
- 116.** Villanueva Espinoza, M.R. 2002. Las Redes Neuronales Artificiales y su Importancia como Herramientas en la Toma de Decisiones. Tesis de Licenciatura, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Académico Profesional de Investigación Operativa, Lima Perú. http://sisbib.unmsm.edu.pe/bibvirtualdata/tesis/basic/Villanueva_EM/enPDF/T_completo.pdf (descargado 01 de enero de 2013).
- 117.** Von der Malsburg, C. 1973. Self-Organization of Orientation Sensitive Cells in the Striate Cortex. *Kybernetik* 14 (2): 85-100. <http://dx.doi.org/10.1007/BF00288907>.
- 118.** Wang, X. Method of Steepest Descent and its Applications. <http://sces.phys.utk.edu/~moreo/mm08/XuWangP571.pdf> (descargado 01 de enero de 2013).
- 119.** Willshaw, D. J. y von der Malsburg, C. 1976. How Patterned Neural Connections Can Be Set Up by Self-Organization. *Proceedings of the Royal Society of London. Serie B, Biological Science* 194 (1117): 431-445. <http://fias.uni-frankfurt.de/~triesch/courses/260theorNS/WillshawVonDerMalsburg1973.pdf> (descargado 01 de enero de 2013).
- 120.** Wolff, M. y Pelissier-Combescure, J. 1982. FACIOLOG – Automatic Electrofacies Determination. SPWLA 23rd Annual Logging Symposium, paper 1982 FF.
- 121.** Wong, P.M., Jian, F.X. y Taggart, I.J. 1995. A Critical Comparison of Neural Networks and Discriminant Analysis in Lithofacies, Porosity and Permeability Predictions. *Journal of Petroleum Geology* 18 (2): 191-206. <http://dx.doi.org/10.1111/j.1747-5457.1995.tb00897.x>.
- 122.** Wyllie, M.R.J., Gregory, A.R. y Gardner, L.W. 1956. Elastic Wave Velocities in Heterogeneous and Porous Media. *Geophysics* 21 (1): 41-70. <http://dx.doi.org/10.1190/1.1438217>.
- 123.** Wyllie, M.R.J., Gregory, A. R., and Gardner, L. W., 1958. An experimental investigation of factors affecting elastic wave velocities in porous media. *Geophysics*, 23(3):459-493.
- 124.** Yu, D., Hu, Q. y Bao, W. 2003. Combining Multiple Neural Networks for Classification Based on Rough Set Reduction. Proc. IEEE Int. Neural Networks and Signal Processing Conf., Nanjing, China, Vol. 1, 543-548. <http://dx.doi.org/10.1109/ICNNSP.2003.1279331>.

-
- 125.** Zapata Garrido, L.A. y Díaz Mojica, H.F. 2007. Predicción del Tipo de Cambio Peso-Dólar Utilizando Redes Neuronales Artificiales. Tesis de Maestría, Fundación Universidad del Norte, Facultad de Ciencias Administrativas, Barranquilla, Colombia. <http://manglar.uninorte.edu.co/jspui/bitstream/10584/143/1/72270588.pdf> (descargado 01 de enero de 2013).
- 126.** Zhang Y., Salisch H.A. y McPherson J.G. 1999. Application of Neural Networks to Identify Lithofacies from Well Logs. *Exploration Geophysics* 30 (2): 45-49. <http://dx.doi.org/10.1071/EG999045>.

Notas: Universidad de Guadalajara (U. de G.) "Tutorial de Redes Neuronales", Cursos Posgrado, 2012, México, Tomado de la página Web de la Universidad Tecnológica de Pereira, Pereira Colombia.