



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA
INGENIERÍA EN EXPLORACIÓN Y EXPLOTACIÓN DE RECURSOS NATURALES
YACIMIENTOS

**“Modelo Neuronal para la Estimación de Valores de
Permeabilidad a Partir de Registros de Pozo y Análisis de
Núcleos”**

TESIS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN INGENIERÍA

PRESENTA:

OMAR ALEJANDRO ARANA HERNÁNDEZ

TUTORA

DRA. SILVIA RAQUEL GARCÍA BENÍTEZ

INSTITUTO DE INGENIERÍA-UNAM

CDMX,

SEPTIEMBRE 2021



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

JURADO ASIGNADO:

Presidente: Dr. Oscar C. Valdiviezo Mijangos.
Secretario: M.C. David Escobedo Zenil.
1^{er}. Vocal: Dra. Silvia Raquel García Benítez.
2^{do}. Vocal: Dr. Néstor Martínez Romero.
3^{er}. Vocal: Dr. Erick Emanuel Luna Rojero.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
División de Estudios de Posgrado. Facultad de Ingeniería.
Ciudad Universitaria. Ciudad de México, México.

TUTOR DE TESIS:

Dra. Silvia Raquel García Benítez



FIRMA

Agradecimientos

A mis padres Lilia Hernández Altamirano y Vicente Arana Arce y a mis hermanos Eduardo y Mariana por siempre estar para apoyarme durante este proceso, durante los buenos y malos momentos, especialmente a lo largo de este último año. Lo que soy y lo que tengo es gracias a ustedes.

A mi tutora la Doctora Silvia Raquel García Benítez por todo el apoyo y enseñanzas que me ha brindado a lo largo de dos años, por ser un ejemplo de constancia y perseverancia y por su compromiso con los estudiantes y con este país.

A mis amigos Henrique y Alejandro por formar parte de esta experiencia que fue la maestría y por todos los momentos que compartimos a lo largo de la misma.

A mis primos Ethiel y Mónica por su apoyo durante las etapas iniciales de este trabajo de investigación.

A los sinodales, por su tiempo para la revisión de este trabajo y comentarios y comentarios sobre el mismo,

A mi alma máter, la Universidad Nacional Autónoma de México, por permitirme formar parte de esta sobresaliente casa de estudios y darme la oportunidad de profundizar mis estudios y conocimientos.

Memento Mori

Contenido

AGRADECIMIENTOS	III
CONTENIDO	IV
RESUMEN	1
ABSTRACT	2
INTRODUCCIÓN	3
ANTECEDENTES	5
CAPÍTULO 1: PERMEABILIDAD	11
1.1 EXPERIMENTO DE DARCY.....	12
1.2 FACTORES QUE AFECTAN LA PERMEABILIDAD DE LA ROCA.....	16
1.3 MEDICIÓN DE LA PERMEABILIDAD.....	21
1.3.1 Pruebas Transitorias de Presión.	26
1.3.2 Modelos Semi-empíricos.....	31
1.3.3 Medición de la Permeabilidad en Laboratorio.....	42
CAPÍTULO 2: CIENCIA DE DATOS	49
2.1 HISTOGRAMAS.....	52
2.2 DIAGRAMA DE CAJAS Y BIGOTES.....	53
2.3 PRUEBA GAMMA.	54
2.4 VALIDACIÓN CRUZADA (K-FOLD CROSS-VALIDATION).....	55
2.5 ÁRBOLES DE DECISIÓN.....	56
2.5.1 Elementos de un Árbol.....	56
2.5.2 Función de Impureza del Nodo.....	57
2.5.3 Poda del árbol	58
2.5.4 Regresión logística o Discriminante Logístico.....	60
2.5.5 Regresión Logística Multinomial.....	61
2.5.6 Algoritmo M5P para Árboles de Regresión.....	61
2.6 ANÁLISIS FACTORIAL.....	62
2.6.1 Modelo de Análisis Factorial	66
CAPÍTULO 3: INTELIGENCIA ARTIFICIAL (IA)	77
3.1 REDES NEURONALES ARTIFICIALES (RNA's).....	83
3.2 ELEMENTOS BÁSICOS DE UNA RED NEURONAL.	85
3.3 VENTAJAS DE LAS REDES NEURONALES.....	86
3.4 TOPOLOGÍAS PRINCIPALES DE LAS REDES NEURONALES.....	88
3.4.1 Función de Entrada.....	90
3.4.2 Función de Activación	91

3.4.3 <i>Función de Salida</i>	93
3.4.4 <i>Mecanismos de Aprendizaje</i>	94
CAPÍTULO 4: RESULTADOS	102
4.1 SOBRE LA KANSAS GEOLOGICAL SURVEY (KGS)	102
4.2 RESUMEN DE CASOS.....	107
4.3 PRIMER ACERCAMIENTO A LA MATRIZ DE DATOS.....	115
4.4 CIENCIA DE DATOS SOBRE LA MATRIZ REDUCIDA.....	132
4.4.1 <i>Resultados: Árbol de Regresión</i>	132
4.4.2 <i>Resultados: Análisis Factorial</i>	153
4.4.3 <i>Resultados: Redes Neuronales</i>	163
4.4.4 <i>Resultados: Modelos Semi-empíricos y comparación</i>	184
CAPÍTULO 5: CONCLUSIONES Y RECOMENDACIONES	195
NOMENCLATURA	198
REFERENCIAS	200
APÉNDICE A	206

Lista de Figuras

Capítulo 1

Figura 1.1. Representación Gráfica de la Permeabilidad. Modificado de (Andersen & Klemm, 2014).	11
Figura 1.2. Experimento de Darcy. Modificado de (Sánchez San Roman, 2005).	13
Figura 1.3 Clasificación del Tamaño de Grano. Modificado de (Hallsworth & Knox, 1999). Sección a) Material con buena clasificación en el tamaño de grano que lo conforman. Sección b) Material con una clasificación regular por la presencia de clastos de menor tamaño. Sección c) Aumenta la presencia de clastos de diversos tamaños lo que disminuye su escala de clasificación. Sección d) material pobremente clasificado donde se presenta una gran variedad de tamaños de grano y por lo tanto menor porosidad y permeabilidad.	17
Figura 1.4. Redondez y Esfericidad. Modificado de (Tucker, 1991). Esta escala permite determinar el nivel de esfericidad (de izquierda a derecha) y redondez (de abajo hacia arriba) de los clastos de un material. Es un reflejo de la composición y grado de intemperismo y transporte que éstos han sufrido.	18
Figura 1.5 Efecto de los Factores que Afectan a la Permeabilidad. Modificado de (Ezekwe, 2011). El grado de redondez, esfericidad y clasificación de un material afectan el comportamiento de la permeabilidad. Materiales con una buena clasificación donde además sus clastos tienen un importante grado de redondez y esfericidad, tienden a dar como resultado valores de permeabilidad mayores. Por el contrario, un material con una clasificación pobre y donde los clastos tengan rastros de poco intemperismo y transporte, se verá reflejado en una disminución de su permeabilidad.	19
Figura 1.6. Empaquetamiento de un material Isótropo y Anisótropo. Modificado de (Freeze & Cherry, 1979). La sección a) presenta un material con un acomodo isótropo de los clastos que lo componen, por lo tanto, no hay una dirección preferente de éstos ni del flujo que podría existir a través. La sección b) muestra, por el contrario, un material con una anisotropía que tiene un acomodo de los clastos preferencial horizontalmente.	21
Figura 1.7. Diagrama de Razonamiento de los Procesos Convencionales para la Determinación de la Permeabilidad.	24
Figura 1.8. Diagrama de Razonamiento de los Procedimientos en Laboratorio.	25
Figura 1.9. Diagramas de comportamiento de la presión y el gasto en una prueba de incremento (a.) y decremento (b.). Modificado de (Lee, 1982).	27
Figura 1.10. Gráfico Especializado de Horner.	28
Figura 1.11. Gráfico Especializado para Flujo radial Infinito.	30
Figura 1.12. Partes Internas de un Barril Nucleador Convencional. Tomada de (Pillado Torres, 2016).	34
Figura 1.13. Barril no-Convencional de corte. Tomado de (Pillado Torres, 2016).	36
Figura 1.14. Empacado de núcleos con papel vinitel. Tomado de (Parrado, 2016).	39
Figura 1.15. Permeámetro de Carga Constante. Modificado de (Torsaeter & Abtahi, 2003).	43
Figura 1.16. Gráfico para la Extrapolación de Klinkenberg.	45
Figura 1.17. Gráfico de Resultados de la Permeabilidad Relativa.	48

Capítulo 2

Figura 2.1 Histogramas.....	53
Figura 2.2 Diagrama de Cajas y Bigotes (Boxplot).....	54
Figura 2.3. Árbol de Decisión.....	57
Figura 2.4. Prueba de Esfericidad de Bartlett.....	65

Capítulo 3

Figura 3.1. Modelos de Inteligencia.	83
Figura 3.2. Analogía de Neuronas Biológicas con una Neurona Artificial.	84
Figura 3.3. Esquematización de una Red Neuronal Artificial multicapa. Modificada de (Matich, 2001).	86
Figura 3.4. Red Neuronal Monocapa.	89
Figura 3.5. Función Lineal de Activación. Modificada de (Matich, 2001).	92
Figura 3.6. Función de Activación Sigmoidea. Modificada de (Matich, 2001).	92
Figura 3.7. Función de Activación Hiperbólica. Modificado de (Matich, 2001).	93
Figura 3.8. Esquema de Bloques del Aprendizaje Backpropagation	97
Figura 3.9. Esquematización del algoritmo backpropagation. Tomado de (Hernández Ambrosio, 2017).	98

Capítulo 4

Figura 4.1. División del estado de Kansas en Condados. Tomado de la KGS	104
Figura 4.2. División en municipio (Township) y rango (Range) de cada uno de los condados. Tomado de la KGS	104
Figura 4.3. División del área en secciones. Tomado de la KGS	105
Figura 4.4. Mapa Interactivo de la KGS.	106
Figura 4.5. Localización de los pozos en los condados de Morton y Stanton. Tomada de KGS	108
Figura 4.6. Área del Campo Hugoton Gas. Tomado de KGS	108
Figura 4.7. Distribución de los Pozos en el Estado de Kansas	109
Figura 4.8. Provincias Estructurales y Distribución de Rocas del Mississippico y el Pennsilvánico en Kansas. Tomado de (Goebel, 1966).	109
Figura 4.9. Campos de gas del sistema Morrow y Atoka en Kansas. Modificado de KGS.....	110
Figura 4.10. Columna Estratigráfica del Suroeste de Kansas (Morrow-Atoka). Modificado de KGS.	111
Figura 4.11. Condiciones Deposicionales de la Formación Morrow. Modificado de KGS	112
Figura 4.12. Zonas Productoras Pertencientes a los Grupos Geológicos Maquoketa y Viola. Modificado de KGS.	113
Figura 4.13. Columna Estratigráfica del Grupo Viola y Maquoketa. Modificado de KGS.	114
Figura 4.14. Ubicación del Pozo de Prueba.	115
Figura 4.15. Boxplot (izquierda) e Histograma (derecha) del Registro ILD.	117
Figura 4.16. Boxplot (izquierda) e Histograma (derecha) del Registro GR.	117

Figura 4.17. Boxplot (izquierda) e Histograma (derecha) del Registro NPHI.	118
Figura 4.18. Boxplot (izquierda) e Histograma (derecha) de la Porosidad.....	118
Figura 4.19. Boxplot (izquierda) e Histograma (derecha) de la Saturación de Aceite.	119
Figura 4.20. Boxplot (izquierda) e Histograma (derecha) de la Saturación de Agua.....	119
Figura 4.21. Boxplot (izquierda) e Histograma (derecha) de la Densidad de Grano.	120
Figura 4.22. Boxplot (izquierda) e Histograma (derecha) de la Permeabilidad.	120
Figura 4.23. Densidad de las Principales Rocas Sedimentarias. Modificada de (Wohlenberg,1982).	123
Figura 4.24. Comportamiento de los Atributos del Pozo 1.....	125
Figura 4.25. Comportamiento de los Atributos del Pozo 2.....	126
Figura 4.26. Comportamiento de los Atributos del Pozo 3.....	127
Figura 4.27. Comportamiento de los Atributos del Pozo 4.....	128
Figura 4.28. Comportamiento de los Atributos del Pozo 5.....	129
Figura 4.29. Comportamiento de los Atributos del Pozo 6.....	130
Figura 4.30. Comportamiento de los Atributos del Pozo 7.....	131
Figura 4.31. Resultado del Árbol de Regresión.....	133
Figura 4.32. Procedimiento para el Pozo 1.	136
Figura 4.33. Permeabilidad Medida Contra Evaluada para el Pozo 1 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	137
Figura 4.34. Procedimiento para el Pozo 2.	138
Figura 4.35. Permeabilidad Medida Contra Evaluada para el Pozo 2 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	139
Figura 4.36. Procedimiento para el Pozo 3.	140
Figura 4.37. Permeabilidad Medida Contra Evaluada para el Pozo 3. (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	141
Figura 4.38. Procedimiento para el Pozo 4.	142
Figura 4.39. Permeabilidad Medida Contra Evaluada para el Pozo 4 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	143
Figura 4.40. Procedimiento para el Pozo 5.	144
Figura 4.41. Permeabilidad Medida Contra Evaluada para el Pozo 5 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	145
Figura 4.42. Procedimiento para el Pozo 6.	146
Figura 4.43. Permeabilidad Medida Contra Evaluada para el Pozo 6 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	147
Figura 4.44. Procedimiento para el Pozo 6.	148
Figura 4.45. Permeabilidad Medida Contra Evaluada para el Pozo 7 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	149
Figura 4.46. Gráfico de Valores de Permeabilidad Medidos contra Evaluados para los Siete Pozos (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	150
Figura 4.47. Procedimiento para el Pozo de Prueba.	151
Figura 4.48. Estimación de los Valores de Permeabilidad para el Pozo de Prueba. (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).....	152
Figura 4.49. Matriz de Correlaciones.	153
Figura 4.50. Matriz de Significancia.	154
Figura 4.51. Gráfico de Sedimentación.	154
Figura 4.52. Matriz de Componente Rotado.....	156
Figura 4.53. Grafico del componente 1 en espacio rotado.	157

Figura 4.54. Instancias del Factor 1	158
Figura 4.55. Gráfico del Componente 2 en Espacio Rotado.....	159
Figura 4.56. Instancias del Factor 2	159
Figura 4.57. Gráfico del Componente 3 en Espacio Rotado.....	160
Figura 4.58. Instancias del Factor 3	161
Figura 4.59. Gráfico del Componente 4 en Espacio Rotado.....	162
Figura 4.60. Instancias Factor 4.....	162
Figura 4.61. Calificación de nodos para el modelo 1, (R-Entr: Coeficiente de correlación del entrenamiento, R-Prueba: Coeficiente de correlación de la prueba).	164
Figura 4.62. Resumen del Modelo de Red 1.....	164
Figura 4.63. Importancia de los Atributos de Entrada para el Modelo de Red 1.	165
Figura 4.64. Permeabilidad Medida Contra Evaluada para la Red 1 (8 entradas, 270 nodos, 1 salida).....	166
Figura 4.65. Predicción de Permeabilidad por la Red 1 (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).	166
Figura 4.66. Predicción de Permeabilidad del Modelo 1 en el Pozo de Prueba (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).	167
Figura 4.67. Calificación de nodos para el modelo 2 (R-Entr: Coeficiente de correlación del entrenamiento, R-Prueba: Coeficiente de correlación de la prueba).	168
Figura 4.68. Resumen del Modelo de Red 1.....	168
Figura 4.69. Importancia de los Atributos de Entrada para el Modelo Neuronal 2.	169
Figura 4.70. Permeabilidad Medida Contra Evaluada para la Red 2 (8 entradas, 50 nodos, 1 salida).....	169
Figura 4.71. Predicción de Permeabilidad por la Red 2. (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).	170
Figura 4.72. Predicción de Permeabilidad del Modelo 2 en el Pozo de Prueba. (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).	171
Figura 4.73. Calificación de nodos para el modelo 3. (R-Entr: Coeficiente de correlación del entrenamiento, R-Prueba: Coeficiente de correlación de la prueba).	172
Figura 4.74. Resumen del Modelo de Red 3.....	172
Figura 4.75. Importancia de los Atributos de Entrada para el Modelo Neuronal 3.	173
Figura 4.76. Permeabilidad Medida Contra Evaluada para la Red 3.	173
Figura 4.77. Predicción de Permeabilidad por la Red 3. (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).	174
Figura 4.78. Predicción de Permeabilidad del Modelo 3 en el Pozo de Prueba. (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).	175
Figura 4.79. Permeabilidad Real y Estimada por la Red 3 en el Pozo 1.....	176
Figura 4.80. Permeabilidad Real y Estimada por la Red 3 en el Pozo 2.....	177
Figura 4.81. Permeabilidad Real y Estimada por la Red 3 en el Pozo 3.....	178
Figura 4.82. Permeabilidad Real y Estimada por la Red 4 en el Pozo 4.....	179
Figura 4.83. Permeabilidad Real y Estimada por la Red 3 en el Pozo 5.....	180
Figura 4.84. Permeabilidad Real y Estimada por la Red 3 en el Pozo 6.....	181
Figura 4.85. Permeabilidad Real y Estimada por la Red 3 en el Pozo 7.....	182
Figura 4.86. Topología del Modelo de Red 3.	183
Figura 4.87. Análisis de Resultados de Redes y Análisis Factorial para el Pozo 5.....	183
Figura 4.88. Análisis de Resultados de Redes y Análisis Factorial para el Pozo 6.....	184
Figura 4.89. Comparación de Resultados para el Pozo 1.	186

Figura 4.90. Comparación de Resultados para el Pozo 3.	188
Figura 4.91. Comparación de Resultados para el Pozo 4.	189
Figura 4.92. Comparación de Resultados para el Pozo 6.	190
Figura 4.93. Comparación de Resultados para el Pozo 7.	191
Figura 4.94. Permeabilidad Medida Contra Evaluada para el Modelo de Pape.	192
Figura 4.95. Permeabilidad Medida Contra Evaluada para el Modelo de Timur.	192
Figura 4.96. Permeabilidad Medida Contra Evaluada para el Modelo de Coates.	193
Figura 4.97. Resultado de Permeabilidad en el Pozo Prueba.	194

Lista de Tablas

Capítulo 1

Tabla 1.1. Métodos Semi-empíricos para la Determinación de la Permeabilidad (Mohaghegh, 1997).	31
Tabla 1.2. Sistemas Convencionales para corte de Núcleos. Modificado de (Exploration and Production Department API, 1998).	34
Tabla 1.3. Sistemas de Corte Especiales para Núcleos. Modificado de (Exploration and Production Department API, 1998).	35
Tabla 1.4. Análisis Rutinario y Mediciones Adicionales. Modificado de (Torsaeter & Abtahi, 2003).	40
Tabla 1.5. Análisis Especial de Núcleos. Modificado de (Torsaeter & Abtahi, 2003).	41

Capítulo 4

Tabla 4.1. Rangos de permeabilidad de rocas. Modificada de (Fetter, 2001).	123
Tabla 4.2. Valores de Estadística Básica de la Base de Datos.	124
Tabla 4.3. Resultados de las pruebas de Esfericidad de Bartlett y KMO.	153
Tabla 4.4. Comunalidades del Modelo Factorial.	155
Tabla 4.5. Varianza Total Explicada.	155
Tabla 4.6. Modelos Semi-empíricos Utilizados.	185
Tabla 4.7. Resumen de los Valores de R y R^2 de los Modelos para el Cálculo de la Permeabilidad.	193

Resumen

En este trabajo se presenta una metodología para estimar, con redes neuronales, RNs, la permeabilidad de la roca a partir de datos de registros de pozo (ILD, GR, NPHI) y características básicas obtenidas de análisis de núcleos (Densidad de Grano, Porosidad, Saturación de agua y Saturación de aceite).

La propuesta brinda una solución integral al reto de parametrizar eficientemente masas naturales. Como preprocesamiento se introduce la aplicación de herramientas de la Ciencia de Datos a la base de geo-informaciones con lo que se profundiza en el entendimiento de las relaciones entre variables y se construye una base sólida con la que es posible sostener o descartar hipótesis de comportamiento. Como modelo predictor se usa una red multicapa con alimentación al frente con algoritmo de aprendizaje de retropropagación.

La RN, entrenada directamente con la información de campo y laboratorio, se muestra como una alternativa confiable a los modelos semi-empíricos normalmente utilizados en la industria. La RN predice valores de permeabilidad con aproximación superior, pero, sobre todo, con conocimiento agregado sobre las rutas paramétricas que generan cada valor de salida. El modelo es legible, asequible, eficiente y autoexplicativo, lo que lo hace ideal como herramienta de acercamiento preliminar en la interpretación de entornos naturales complejos.

Abstract

In this work, a methodology to estimate, with neural networks NNs, rock permeability from well log data (ILD, GR, NPHI) and basic characteristics obtained from core analysis (Grain Density, Porosity, Water saturation and Oil saturation), is presented.

This approach provides a comprehensive solution to the challenge of efficiently parameterizing natural masses. As preprocessing, application of Data Science tools to the geo-information database is introduced, which deepens the understanding of the relationships between variables and builds a solid database on which behavioral hypotheses can be supported or discarded. A feed-forward multilayer network with backpropagation as learning algorithm is used as the predictor model.

The RN, trained directly with field and laboratory information, is shown as a reliable alternative to the semi-empirical models used in the industry. This RN predicts permeability values with a superior approximation, but above all, with added knowledge about the parametric routes that generate each output value. The model is legible, affordable, efficient and self-explanatory, making it ideal as a preliminary approach tool in interpreting complex natural environments.

Introducción

Este trabajo de investigación tiene como propósito presentar una metodología para estimar la permeabilidad a partir de datos de porosidad, resistividad y radiactividad del medio (obtenidos de registros petrofísicos de pozo) y características básicas de núcleos. Para lograr el objetivo se utilizarán las ventajas de modelado de las Redes Neuronales (Multicapa, de Alimentación al Frente, con entrenamiento supervisado, aprendizaje con Propagación Hacia Atrás) como recursos computacionales que permiten correlacionar parámetros que conforman bases de datos complejas.

Obtener información suficiente para la caracterización de yacimiento y diseños de acciones sobre él (para los correspondientes modelos) es aún una actividad en constante evolución y que demanda mejoras significativas. Geólogos, petrofísicos e ingenieros requieren de conjuntos de variables para activar modelos o establecer simulaciones al respecto de posibles tasas de producción, reservas y factores de recuperación (ya sea en terminación o estimulación de sitios). Entre estas variables, la permeabilidad es una de las características esenciales de la roca que influye considerablemente en las decisiones sobre el desarrollo de los campos, por lo que la confianza sobre el método para obtenerla es fundamental.

Los métodos tradicionales para medir o calcular la permeabilidad podrían arrojar en algunos escenarios valores poco representativos que inciden negativamente en los costos, los tiempos y la seguridad. Son las muestras de núcleos las que proporcionan detalles valiosos para completar las necesidades de información, sin embargo, pruebas de este tipo normalmente se aplican en sub-regiones muy pequeñas del volumen de interés (Varhaug & Smithson, 2015). Esta investigación intenta colocar un método sencillo en el que se utilicen a las redes neuronales para aprovechar la información de estas pruebas y caracterizaciones simples en la predicción, como primera aproximación, de un valor de permeabilidad confiable.

El modelo neuronal resultante es alimentado con matrices de datos que pasan filtros de ciencia de datos en un esfuerzo por extraer conocimiento valioso sobre los patrones de

comportamiento entre las mediciones y las características de los núcleos. El modelo construido así se presenta como local y confiable.

La historia de este tipo de aplicaciones en la industria petrolera no es reciente. Desde los años 90 que Ali J. K (1994) presentó un panorama general de las aplicaciones directas de las redes neuronales a la industria del petróleo, la atención y los esfuerzos no han cesado. La aportación original de este trabajo se centra en la adición de rutas de análisis numérico para refinar las bases numéricas y lingüísticas, y el incremento del tamaño de las propias bases de entrenamiento y prueba, de forma que la solución y el método puedan ser considerados una buena alternativa a la solución del problema fundamental de la ingeniería de yacimientos: la eficiente caracterización paramétrica del medio.

Este documento se estructura sobre las siguientes secciones:

1. Descripción de la problemática, donde se detalla la justificación, el alcance de la investigación y los objetivos principales.
2. Breve revisión de aspectos teóricos básicos para delimitar el tipo de estudios de permeabilidad en la Industria Petrolera y, por otro lado, los desarrollos relacionados con inteligencia artificial – redes neuronales.
3. Revisión de algunos métodos convencionales para la determinación de la permeabilidad que sirve de base para las secuencias de modelado *inteligente*.
4. Ciencia de Datos aplicada a la matriz de informaciones.
5. Uso de redes neuronales para el cumplimiento del objetivo planteado.

Antecedentes

Se reconoce que la explotación de las herramientas de la inteligencia artificial en la industria del petróleo, sobre todo para desarrollar modelos con redes neuronales, han existido hace varias decenas de años, pero es en los más recientes que se ha verificado un interés absoluto dadas las contundentes pruebas de éxito en su aplicación. Por esto, los antecedentes que se citan a continuación no intentan ser un Estado del Arte de tantos años de esfuerzos sino un resumen de los trabajos que inspiran a quien presenta esta Tesis dada su afinidad teórica y significado sobre el tema.

Las investigaciones sobre la permeabilidad deben tratar con un problema complejo (o muy complejo) que requiere de un manejo matemático dispendioso (y en muchas aristas incierto) y que en gran número de casos podría no generar los mejores resultados. Explotar la aplicación de nuevas tecnologías, como las redes neuronales, para optimizar este proceso tan importante de la industria petrolera resulta un camino natural dadas las características de aprendizaje (aprender a partir de ejemplos), auto organización (crear su propia representación de la información internamente), tolerancia a fallos (responder de manera aceptable aún con información o estructuras dañadas parcialmente) y flexibilidad (manejar cambios no importantes en la información de entrada, como señales con ruido). A continuación, se presentan algunos de los trabajos que, a juicio del autor, representan avances significativos en el aprovechamiento de esta herramienta.

- Güler & Ertekin (1999) estudiaron el desarrollo de un modelo de red neuronal para la estimación de permeabilidades relativas. Este estudio se realizó sobre datos obtenidos netamente de experimentos en laboratorio (15 y 13 fueron usados para el entrenamiento). Cabe resaltar que algunas de las permeabilidades se obtuvieron a través de formulaciones matemáticas convencionales (el modelo usado no es especificado). Uno de los problemas que enfrenaron fue la precaria predicción de permeabilidad relativa al aceite (valores no satisfactorios, **Figura 1**).

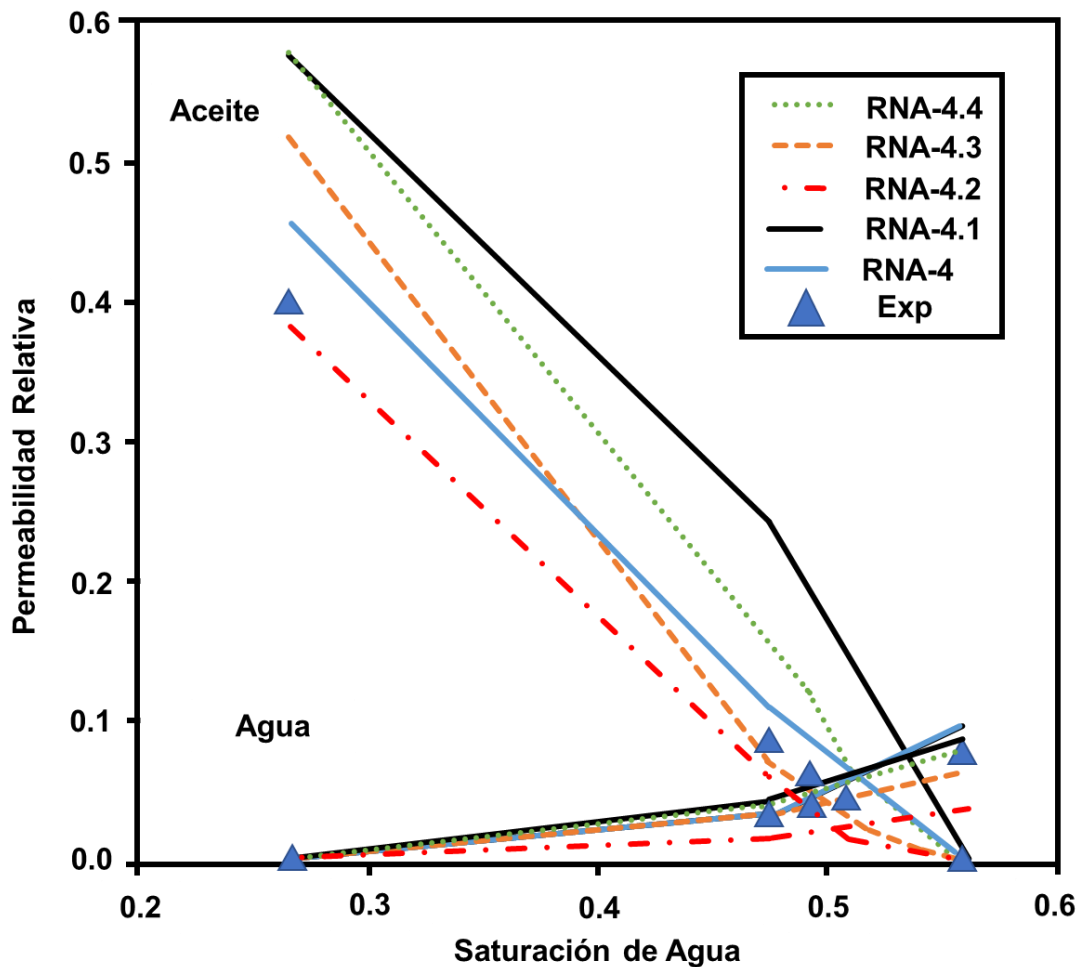


Figura 1. Valores experimentales y predichos de permeabilidad relativa para distintos modelos neuronales, donde varían algunas de los parámetros de entrada presentados ya sea porque se agregaron resultados de experimentos más complejos o debido a que algunos de los ya existentes fueron trabajados para entrar a la red como una transformación (ej. La inversa del parámetro o su cuadrado). Modificado de (Güller & Ertekin, 1999).

- Aminian K. y colaboradores (2003) predijeron el desempeño de un yacimiento a través de la aproximación de valores de permeabilidad y determinación de unidades de flujo haciendo uso de redes neuronales supervisadas y no supervisadas, respectivamente. Se resaltan problemas para identificar los mejores conjuntos de datos en entrenamiento (**Figura 2**) y las dificultades para aproximar valores en las zonas de transición entre unidades de flujo.

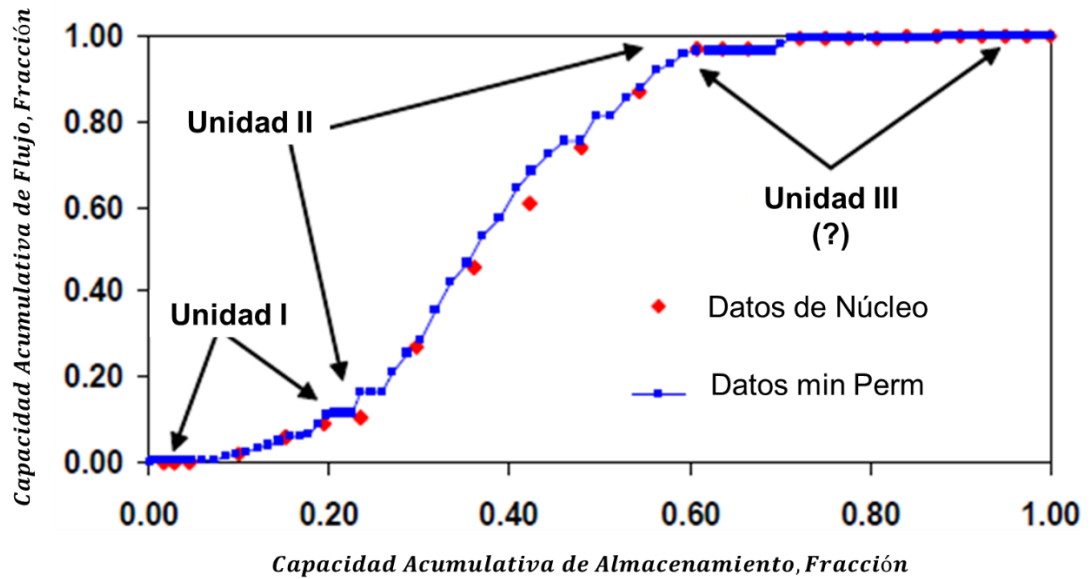


Figura 2. Gráfico de Unidades de Flujo Determinadas por la Metodología Descrita. Tomada de (Aminian et al, 2003).

- En 2005 Singh S. (Singh, 2005) presentó un procedimiento para determinar la permeabilidad a partir de registros de pozo de rayos gamma, neutrón y densidad, en un área de la cuenca de Uinta, Utah. Este trabajo hace gran énfasis en la importancia de la calidad de los datos usados en el entrenamiento. Se concluye que la red neuronal puede manejar correctamente heterogeneidades a pequeña escala. Sin embargo, hay que hacer notar que sus rangos dinámicos son pequeños: GR-0 a 30u, Neutrón-10 a 18u, Densidad-2.3 a 2.87. El entrenamiento de la red se hizo con información de siete pozos y la validación sobre seis pozos que fueron reservados para ello, (coeficientes de correlación medido contra evaluado entre 0.44 y 0.95 se muestran en la **Figura 3**).

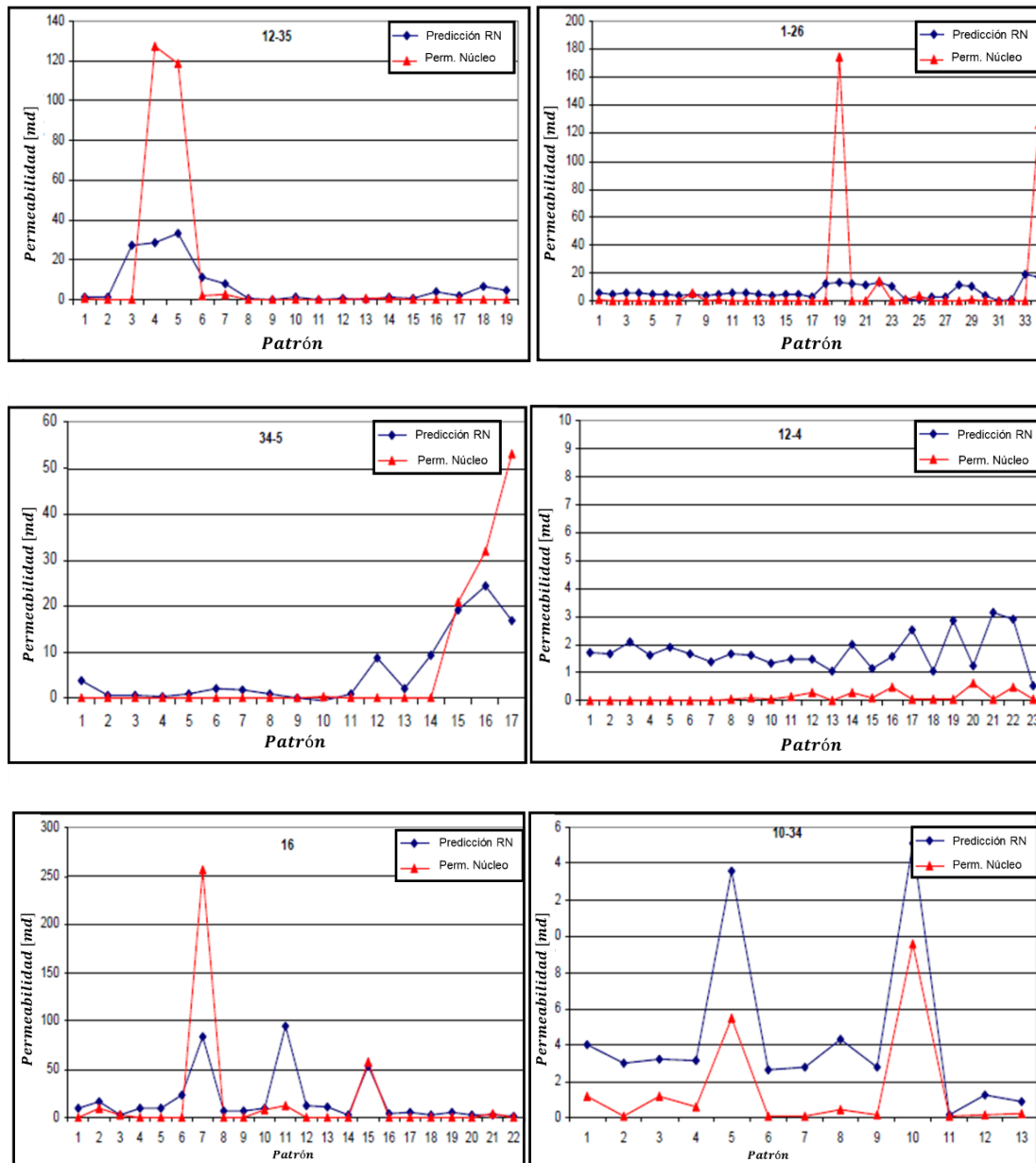


Figura 3. Resultados de Permeabilidad Obtenida de Núcleo Contra Permeabilidad Estimada con una Red Neuronal usando la Metodología Resaltada. Tomado de (Singh, 2005).

- En 2013 Maslennikova (Maslennikova, 2013) desarrolló una red neuronal híbrida para la determinación de la permeabilidad utilizando técnicas de agrupación (clustering) sobre la matriz de atributos. Teniendo como entradas la porosidad, densidad, resistividad y saturación de agua, todas obtenidas de registros de pozo, generó un modelo neuronal por cada clúster definido (pre-procesamiento de datos

con segmentación automática). Llegó a la conclusión de que este enfoque puede disminuir en gran medida el error del modelo predictor de permeabilidad, pues existe menor dispersión entre los atributos que pertenecen a un mismo cluster que aquella existente considerando a todos los atributos en una (única) matriz de datos (Figura 4).

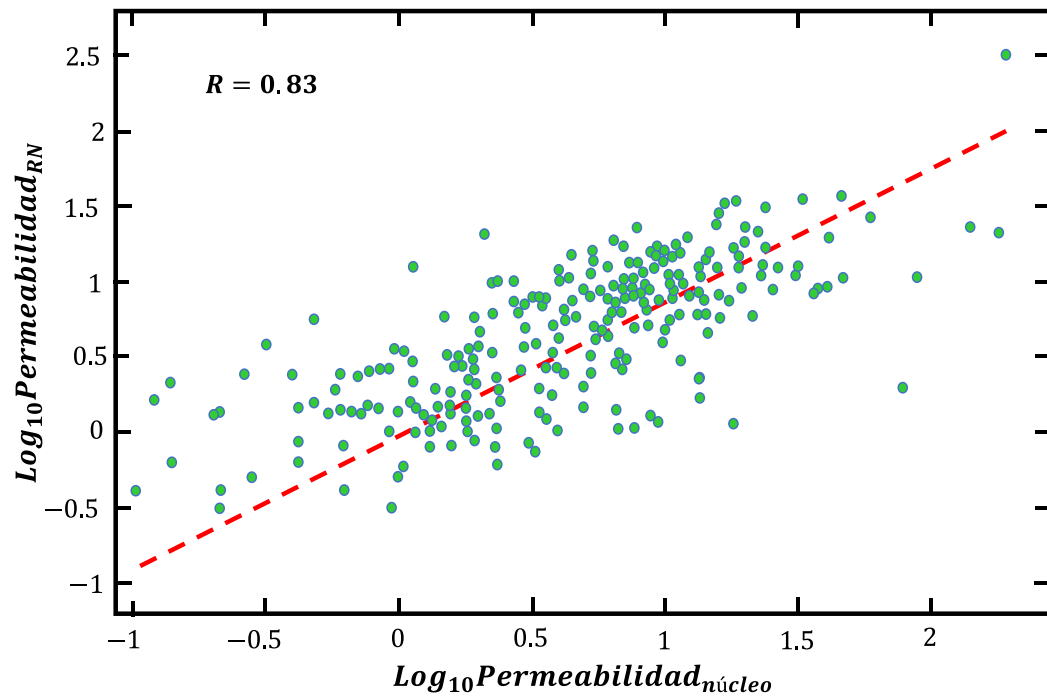


Figura 4. Resultados del Modelo Neuronal Híbrido. Modificado de (Maslennikova, 2013).

- En el año 2014 Kohli y Arora (Kohli & Arora, 2014) presentaron una metodología para la aproximación de valores de permeabilidad a partir de datos de tres registros de pozos y los resultados del análisis de núcleos. No se aclara si los datos (alrededor de 60 instancias) hayan tenido algún tipo de pre-procesamiento. Se concluye que el potencial de estos modelos para la determinación de propiedades en posiciones del medio donde no se cuentan con informaciones absolutas, es muy alto. Dos de los tres pozos se usaron para el entrenamiento y con el último se validó el modelo (**Figura 5**). La estimación de la permeabilidad, con algunos valores negativos, no es del todo aceptable.

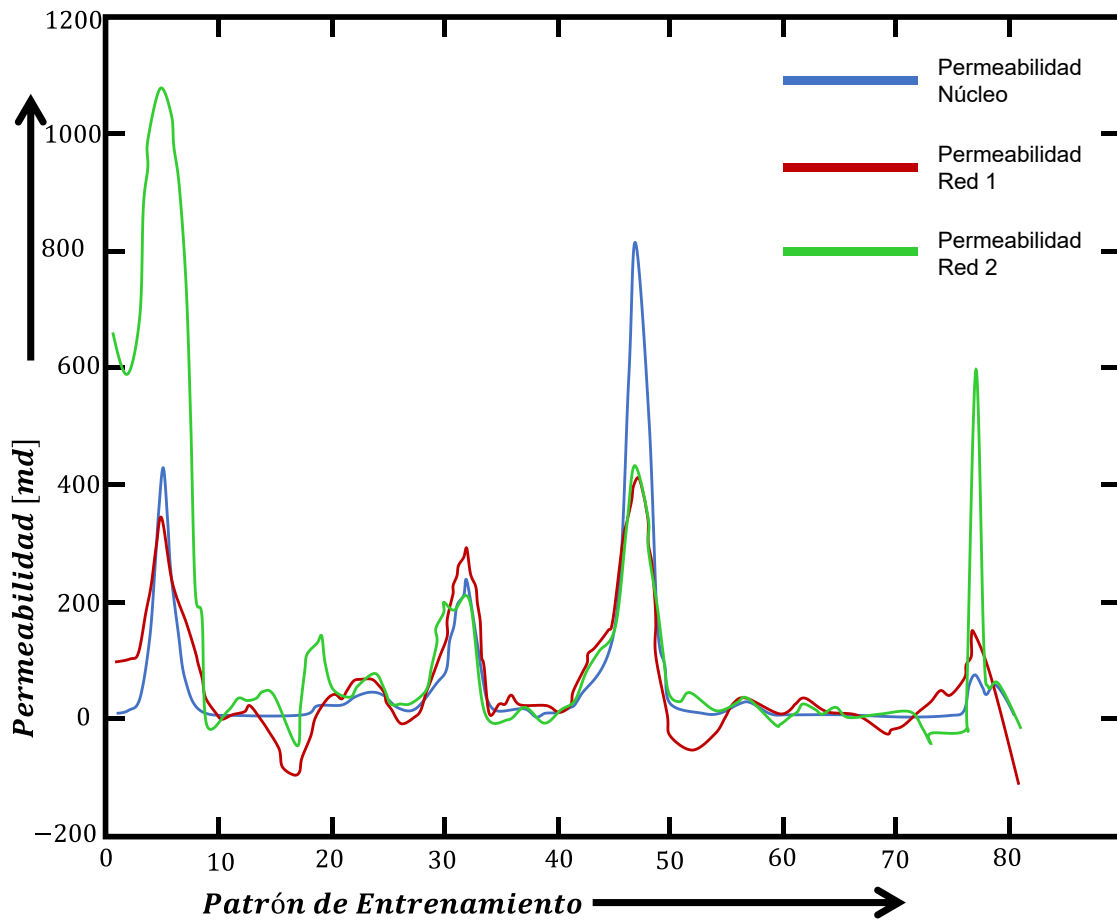


Figura 5. Resultados de permeabilidad Estimada por el Modelo Neuronal contra los valores obtenidos del análisis de Núcleos. Modificado de (Kohli & Arora, 2014).

Capítulo 1: Permeabilidad

A continuación, se presenta un panorama general sobre la permeabilidad, los procesos geológicos deposicionales y diagenéticos que la afectan y modifican, así como las metodologías para su medición o determinación tanto en campo como en laboratorio. Se resalta, además, la importancia que tiene esta propiedad dentro de la Industria Petrolera, especialmente en el área de yacimientos.

La permeabilidad se define como la capacidad de un material poroso para dejar pasar a través de éste un fluido. Es una propiedad intrínseca que controla la facilidad con la cual los fluidos se mueven en el medio en yacimientos de hidrocarburos, acuíferos, filtros y paquetes de grava. Depende de factores geométricos, de tamaño e interconexión de los poros y de las características capilares y de fracturamiento (Andersen & Klemin, 2014) **(Figura 1.1)**

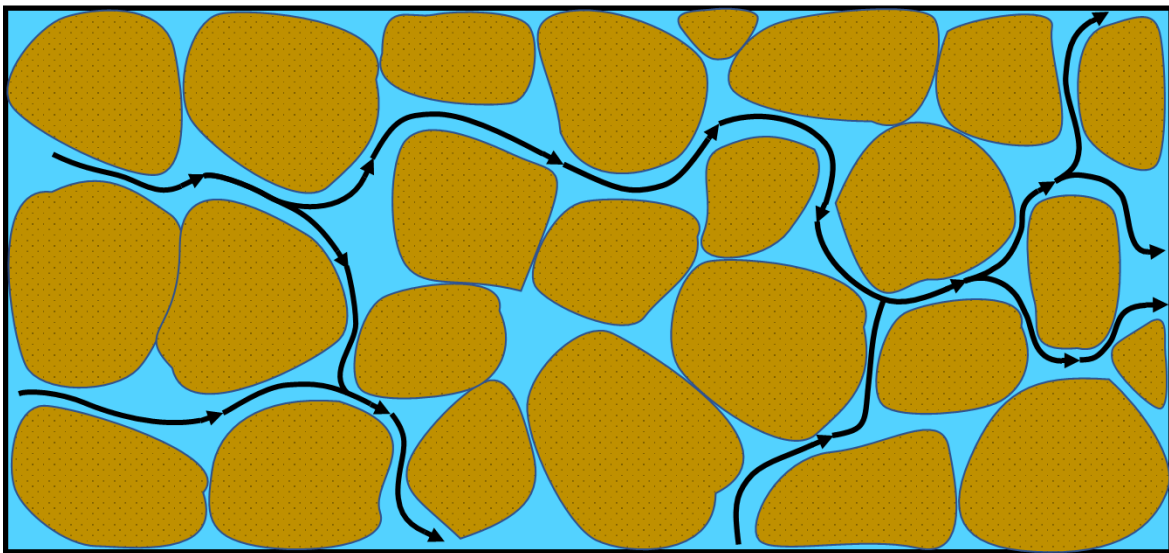


Figura 1.1. Representación Gráfica de la Permeabilidad. Modificado de (Andersen & Klemin, 2014).

1.1 Experimento de Darcy

El primer trabajo sobre permeabilidad fue conducido por Darcy (1856) y su interés era el estudio del flujo de agua a través de medios porosos pues en su ciudad (Dijon, Francia) se utilizaban filtros de arena para depurar el agua (Pinder & Celia, 2006).

El experimento de Darcy consiste básicamente en hacer circular agua a través de un recipiente de sección constante el cual contiene un empaquetamiento de algún material poroso, en este caso arenas. Uno de sus extremos es conectado a un depósito con cierta elevación, mientras que en el otro extremo el flujo de salida es regulado mediante un grifo a gasto constante. La altura de la columna de agua debe ser medida en un mínimo de dos puntos a lo largo del recipiente (**Figura 1.2**). Al repetir el experimento con varios materiales porosos (como filtro del recipiente) Darcy dedujo que el gasto que atraviesa el medio poroso es directamente proporcional a la sección y al gradiente hidráulico, y que, además, la constante de proporcionalidad, conocida como permeabilidad (k) en la ecuación 1.1, era característica de cada material que ocupa el recipiente (Darcy, 1856).

$$q = k * A_{sección} * \frac{\Delta h}{\Delta l}, \dots\dots\dots (1.1)$$

donde

$q = \text{Gasto } [L^3 / t]$.

$k = \text{constante de proporcionalidad}$.

$A_{sección} = \text{Área de la sección del recipiente } [L^2]$.

$\frac{\Delta h}{\Delta l} = \text{Gradiente hidráulico } [1]$.

Cabe resaltar que el valor de k es constante sólo en un medio homogéneo.

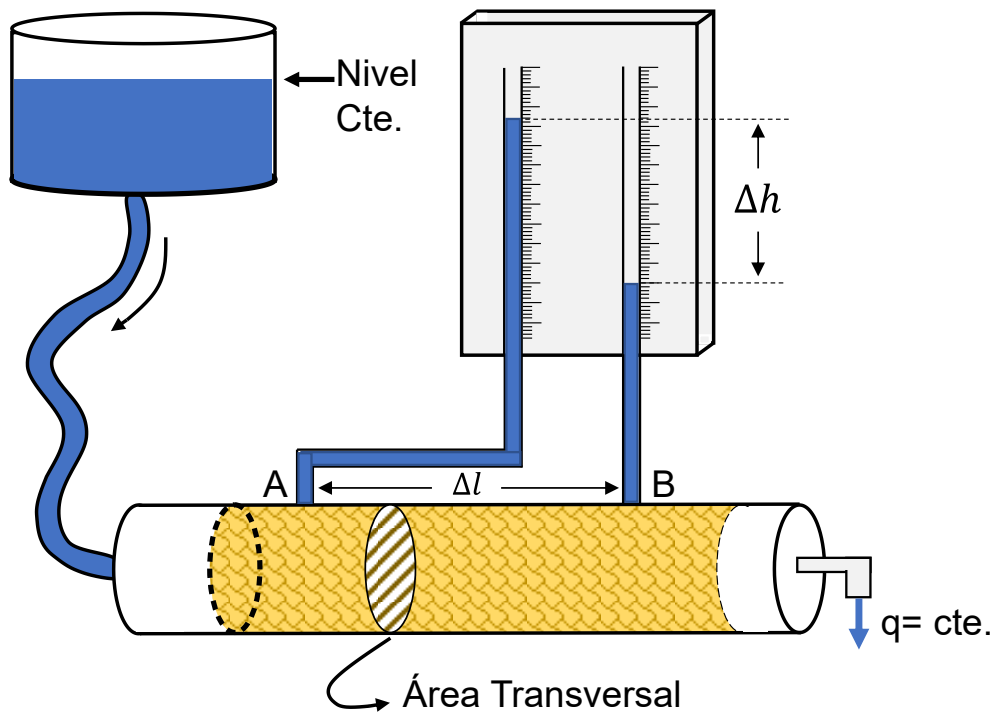


Figura 1.2. Experimento de Darcy. Modificado de (Sánchez San Roman, 2005).

Finalmente, la expresión de la Ley de Darcy es la siguiente:

$$q = -k \left(\frac{dh}{dl} \right). \dots\dots\dots (1.2)$$

En la ecuación 1.2 el signo negativo se debe a que el flujo del fluido en estudio se da de las zonas de mayor energía a las de menor energía.

En años posteriores el trabajo de Darcy fue retomado por Muskat y Botset (Muskat, 1937; Botset, 1933) quienes publicaron una formulación superior que es ampliamente usada en la industria hoy en día (ecuación 1.3):

$$q = - \frac{k * A \Delta P}{\mu * L}, \dots\dots\dots (1.3)$$

donde

$q = \text{Gasto [bpd]}$.

$k = \text{Permeabilidad [md]}$.

$\Delta P = \text{Caída de presión a lo largo de la muestra [psi]}$.

$A = \text{Área transversal de la muestra [in]}$.

$L = \text{Longitud de la muestra [in]}$.

$\mu = \text{Viscosidad del Fluido [cp]}$.

La unidad de permeabilidad más utilizadas dentro de la industria es el Darcy. Este se define como la permeabilidad que permite a un fluido de 1 centipoise (cP) de viscosidad fluir a una velocidad de 1 cm/s para una caída de presión de 1 atm/cm . Debido a que la mayoría de los yacimientos tienen permeabilidades mucho menores a un Darcy, el milidarcy (md) es comúnmente empleado (Selley, 1998).

La Ley de Darcy es únicamente válida si:

- no existen reacciones químicas entre el fluido y la roca,
- la saturación del medio es del 100% de un solo fluido,
- existe flujo laminar, y
- los sistemas son del tipo poro uniforme (para medios de doble porosidad donde existan relaciones complejas entre sus partes la Ley de Darcy no puede representarlas de forma exacta).

Tipos de Permeabilidad

Si el medio poroso se encuentra saturado completamente por un fluido, la permeabilidad medida en ese medio se referirá a la permeabilidad absoluta (K o K_a). Cuando el medio poroso se encuentra ocupado por más de un fluido, la propiedad medida es la permeabilidad efectiva (K_{ef}) de dicho medio a uno de los fluidos en particular. Finalmente, la permeabilidad relativa (K_{rf}) se define como la relación de la permeabilidad efectiva entre la permeabilidad absoluta de la roca (Selley, 1998).

Esta propiedad no debe confundirse con la movilidad ni con la conductividad hidráulica. La movilidad se define como la permeabilidad relativa de un fluido a un medio dividida

por la viscosidad dinámica del fluido que viaja a través de dicho medio, tiene unidades de uno entre centipoise (Andersen & Klemin, 2014), y se define como:

$$m_o = \frac{k_{rf}}{\mu_f}, \quad \dots\dots\dots (1.4)$$

donde

$m_o =$ *movilidad.*

$k_{rf} =$ *permeabilidad relativa a un fluido.*

$\mu_f =$ *viscosidad dinámica del fluido.*

La conductividad hidráulica, o transmisividad, es la descarga o velocidad efectiva a través del medio y equivale al volumen de fluido que atraviesa una sección transversal durante un intervalo de tiempo, dividido por la sección transversal, tiene dimensiones de longitud entre tiempo, y se expresa como:

$$CH = \frac{q}{A}, \quad \left[\frac{L}{T} \right] \quad \dots\dots\dots (1.5)$$

donde

$CH =$ *Conductividad Hidráulica.*

$q =$ *Gasto de fluido.*

$A =$ *Área transversal al gasto.*

La movilidad y la conductividad hidráulica son características colectivas que combinan las propiedades del fluido con las del medio poroso (Pinder & Celia, 2006).

1.2 Factores que Afectan la Permeabilidad de la Roca

En una gran cantidad de materiales es posible considerar que la permeabilidad es directamente proporcional a la porosidad (fracción del volumen total del material ocupada por poros o vacíos). Sin embargo, esta no es una regla absoluta. Los factores texturales y geológicos determinan la magnitud de la permeabilidad mediante el incremento o la reducción de la sección transversal del espacio poroso abierto y son independientes del tipo de fluido (Andersen & Klemin, 2014).

Los materiales formados a partir de estructuras apiladas de esferas sólidas idénticas sean balas de cañón, canicas o cojinetes, poseen las mismas porosidades. Sin embargo, las secciones transversales de los poros difieren significativamente; por consiguiente, las permeabilidades de estas estructuras también difieren. La permeabilidad de las rocas compuestas por granos grandes, o gruesos, será mayor que la de los granos pequeños o finos. La permeabilidad también es influenciada por la forma de los granos. Las medidas de la forma de los granos son la esfericidad, la redondez y la rugosidad, por lo tanto, es importante definir y entender dichos factores con la finalidad de construir una base para entender el comportamiento y distribución de la permeabilidad en un medio rocoso.

La Clasificación de los granos de roca (**Figura 1.3**) es el rango de tamaños existentes en los materiales sedimentarios. Los materiales con buena clasificación poseen granos del mismo tamaño (Figura 1.3a) en tanto que los materiales pobremente clasificados poseen granos de tamaños diversos (1.3d). La permeabilidad se reduce a medida que el grado de clasificación varía de buena a pobremente clasificado (Figuras 1.3b y 1.3c) ya que los granos de menor tamaño pueden rellenar los espacios existentes entre los granos más grandes (Hallsworth & Knox, 1999).

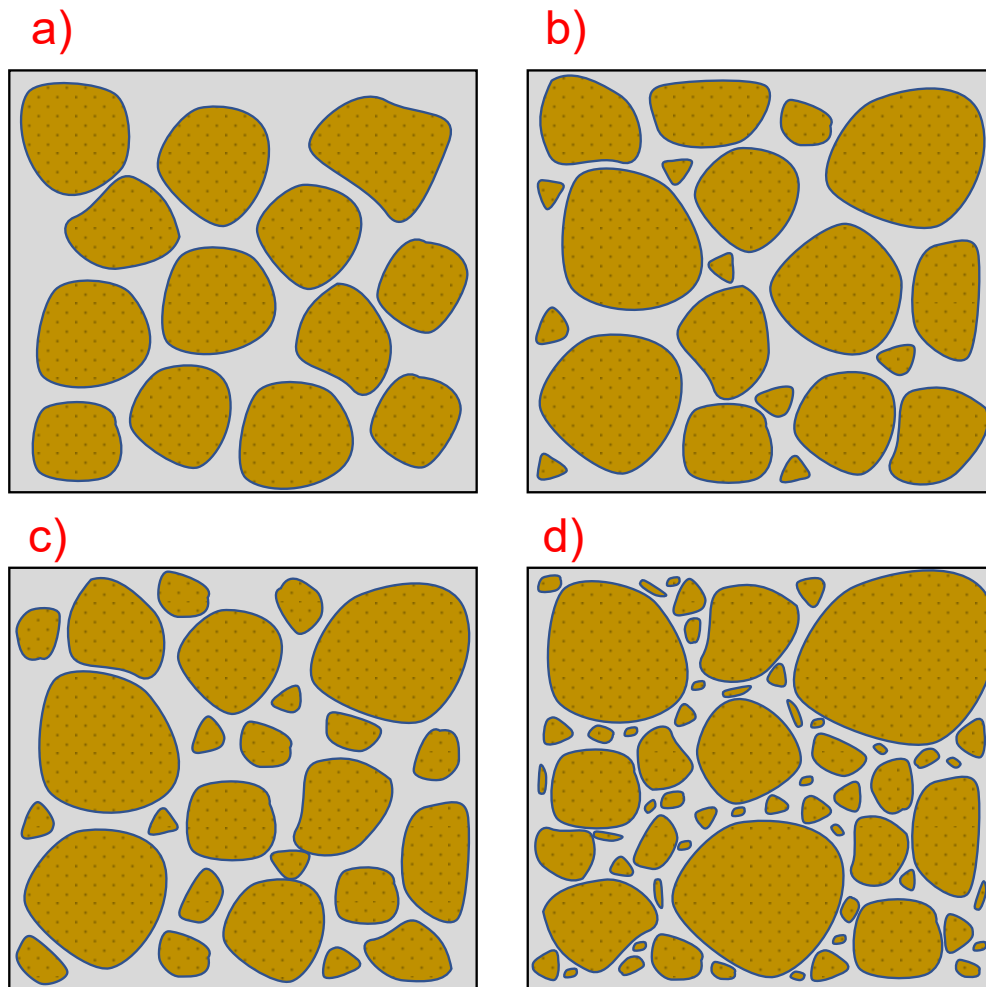


Figura 1.3 Clasificación del Tamaño de Grano. Modificado de (Hallsworth & Knox, 1999). Sección a) Material con buena clasificación en el tamaño de grano que lo conforman. Sección b) Material con una clasificación regular por la presencia de clastos de menor tamaño. Sección c) Aumenta la presencia de clastos de diversos tamaños lo que disminuye su escala de clasificación. Sección d) material pobremente clasificado donde se presenta una gran variedad de tamaños de grano y por lo tanto menor porosidad y permeabilidad.

La *Esfericidad* se refiere a en qué grado la forma de un clasto se aproxima a la de una esfera mientras que la *redondez* se refiere al grado de erosión de los cantos de la partícula, es decir, se liga a la agudez de los bordes y esquinas (varía entre muy anguloso a muy redondeado) (Tucker, 1991). En la **Figura 1.4** se muestra un clasto de alta esfericidad que va cambiando su redondez hasta un clasto con esfericidad baja que, de igual forma, va cambiando su redondez conforme baja en la escala.

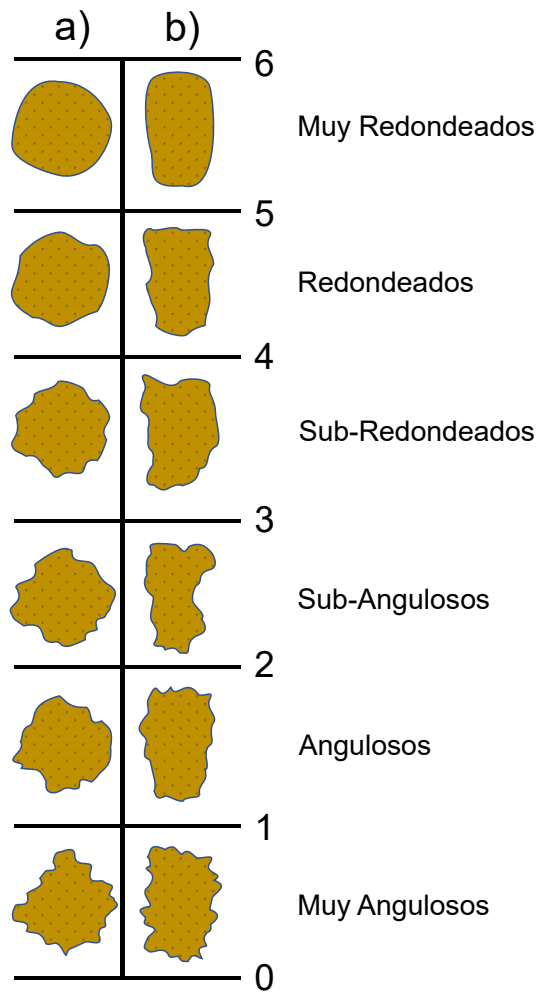


Figura 1.4. Redondez y Esfericidad. Modificado de (Tucker, 1991). Esta escala permite determinar el nivel de esfericidad (de izquierda a derecha) y redondez (de abajo hacia arriba) de los clastos de un material. Es un reflejo de la composición y grado de intemperismo y transporte que éstos han sufrido.

La rugosidad es el grado de textura de grano que afecta el empaquetamiento; es decir, su disposición 3D en el medio. La variabilidad de la forma de los granos puede impedir que se alcance el empaquetamiento más compacto posible, lo que incide en la permeabilidad. A medida que se incrementa el grado de empaquetamiento, pasando de no consolidado a compacto, un grano se pone en contacto con un número cada vez mayor de granos adyacentes. En consecuencia, los espacios existentes entre los granos y las secciones transversales abiertas al flujo se reducen, lo que se traduce en menor permeabilidad (Tucker, 1991).

En la **Figura 1.5** se muestran los valores más altos posibles teniendo clastos bien clasificados con una alta redondez y esfericidad, hasta aquellos con valores de permeabilidad muy baja, pudiendo existir en ambos casos una diferencia considerable entre la permeabilidad horizontal y la permeabilidad vertical.

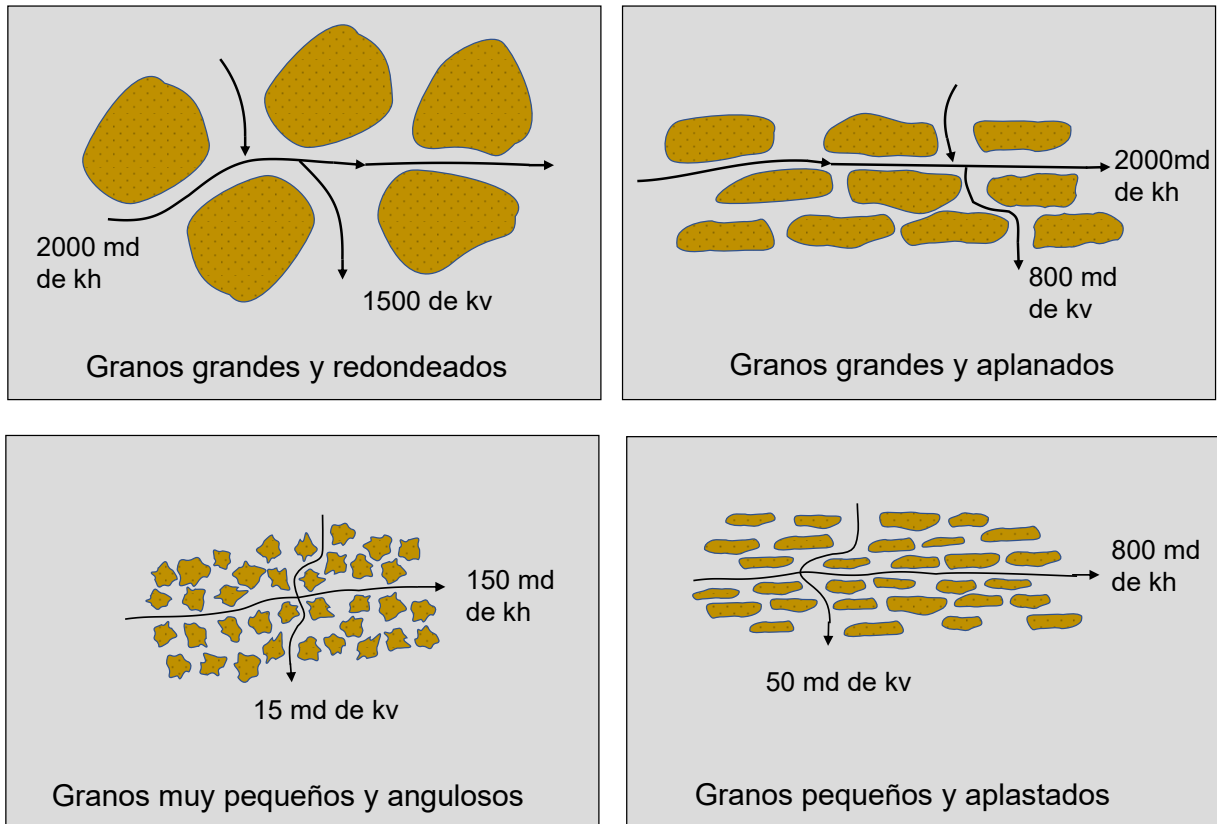


Figura 1.5 Efecto de los Factores que Afectan a la Permeabilidad. Modificado de (Ezekwe, 2011). El grado de redondez, esfericidad y clasificación de un material afectan el comportamiento de la permeabilidad. Materiales con una buena clasificación donde además sus clastos tienen un importante grado de redondez y esfericidad, tienden a dar como resultado valores de permeabilidad mayores. Por el contrario, un material con una clasificación pobre y donde los clastos tengan rastros de poco intemperismo y transporte, se verá reflejado en una disminución de su permeabilidad.

Otro punto de gran importancia es el *grado de litificación* de la roca, en otras palabras, que tan cementados y consolidados se encuentran los granos que la conforman. Esta propiedad es inherente de los procesos de formación de la roca y se ve especialmente modificada por la naturaleza del ambiente de depósito, sin embargo, puede ser alterada

por procesos diagenéticos, los cuales tienen lugar una vez sea ha formado y consolidado la roca (Andersen & Klemin, 2014).

Diagénesis son todos aquellos cambios físicos y químicos que ocurren en los sedimentos o en las rocas sedimentarias después del depósito que dan como resultado una alteración textural y mineralógica (Bubenick et al, 1938). La disolución, la dolomitización, el fracturamiento y otros procesos que alteran las rocas, además de poder generar una porosidad adicional o secundaria, pueden incrementar o disminuir la permeabilidad.

La reducción del volumen del sedimento (compactación), ocasionada principalmente por las fuerzas verticales ejercidas por una capa de recubrimiento creciente, o por fluidos intersticiales, deforma y reorienta los granos, afectando así la porosidad y permeabilidad de la roca a medida que el sedimento es sepultado. Las presiones de los fluidos también afectan la permeabilidad; un incremento de la presión del fluido abre los poros, en tanto que una reducción produce su cierre (Bubenick et al, 1938). La precipitación del cemento (cementación) entre los granos minerales o los granos de las rocas reduce la permeabilidad, éste puede ocurrir simultáneamente a la sedimentación, o bien, el cemento puede ser introducido en un tiempo posterior. Los minerales de arcilla pueden formar cristales (recristalización) que revisten las paredes de los poros o crecer como fibras y láminas que obturan el volumen poroso, debido a un incremento de esfuerzos y otras influencias. Las arcillas intersticiales autógenas, que son las arcillas que se desarrollan entre los granos, pueden rellenar el espacio poroso y reducir la permeabilidad. Las arcillas alogénicas, que son las arcillas que han sido transportadas hacia el interior de los poros, pueden obturarlos (Andersen & Klemin, 2014).

La mayoría de las rocas exhiben cierta anisotropía de permeabilidad, que se entiende como la variación de la propiedad respecto al cambio de dirección en el espacio. La esfericidad de los granos y la presencia de fracturas son factores que afectan la direccionalidad de la permeabilidad. Como se aprecia en la **Figura 1.6** los granos esféricos forman empaquetamientos isotrópicos que permiten que el fluido fluya bien en todas las direcciones mientras que los granos aplastados (achatados) y ovalados (alargados) –que tienden a orientarse en sentido horizontal, paralelos unos con respecto a otros formando capas- afectan la facilidad del flujo de fluido.

La permeabilidad anisotrópica es mayor cuando los fluidos fluyen en sentido paralelo a una capa que cuando lo hacen en sentido perpendicular a la misma. Los fluidos se mueven con más facilidad a través de las fracturas abiertas que entre los granos. Si las fracturas exhiben una alineación preferencial, la permeabilidad alcanza un valor máximo en sentido paralelo a esta dirección y es anisotrópica. Como consecuencia de los factores texturales y geológicos que inciden en la permeabilidad, el trayecto que recorre el fluido a través de la roca puede ser más largo, con muchos giros y curvas, que la distancia lineal directa entre el punto inicial y el punto final (Fetter C. , 2001).

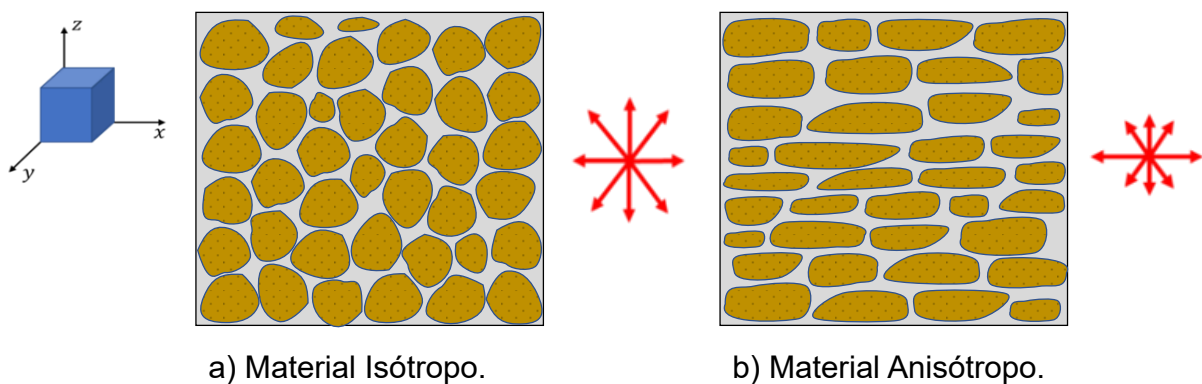


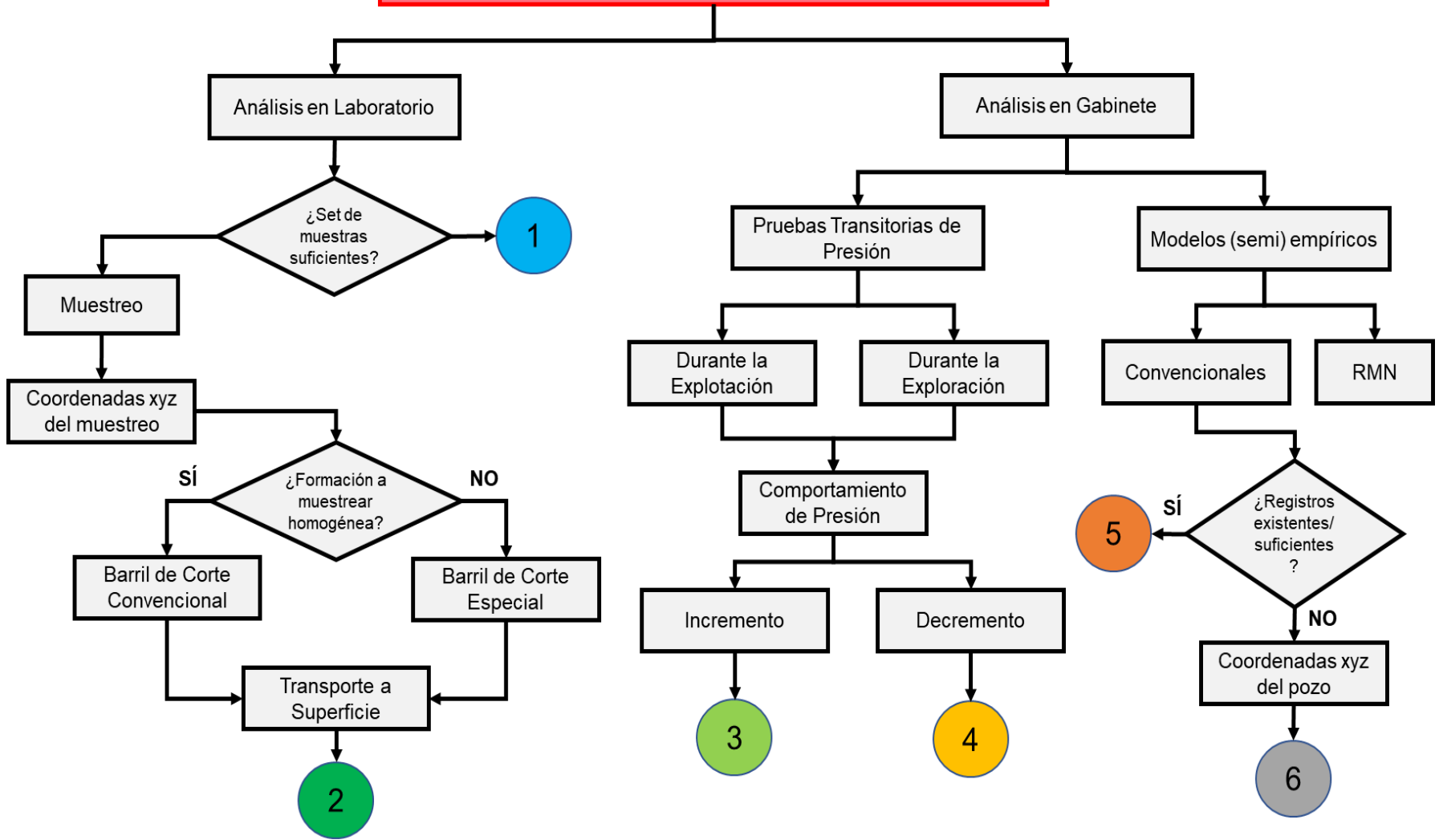
Figura 1.6. Empaquetamiento de un material Isótropo y Anisótropo. Modificado de (Freeze & Cherry, 1979). La sección a) presenta un material con un acomodo isótropo de los clastos que lo componen, por lo tanto, no hay una dirección preferente de éstos ni del flujo que podría existir a través. La sección b) muestra, por el contrario, un material con una anisotropía que tiene un acomodo de los clastos preferencial horizontalmente.

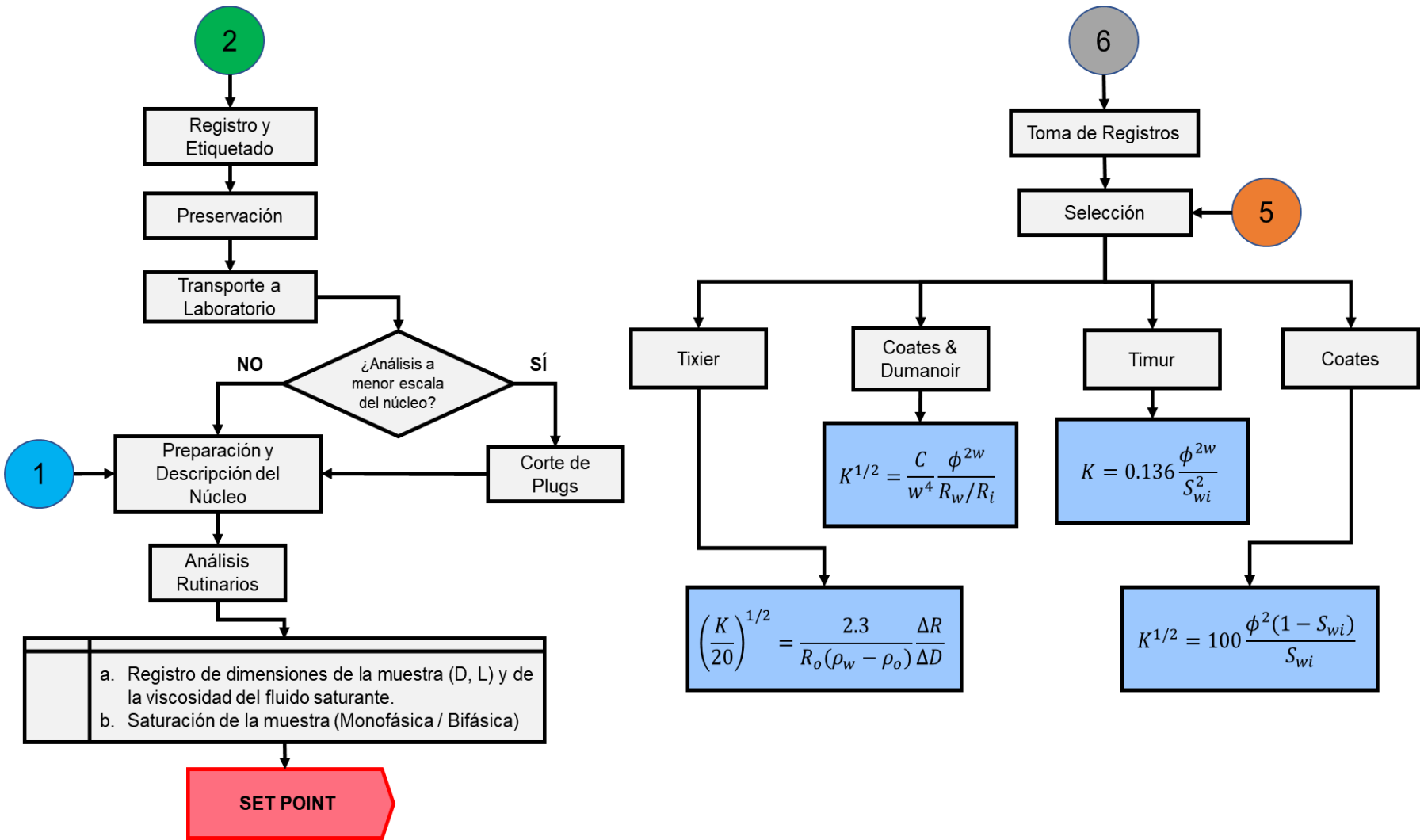
1.3 Medición de la Permeabilidad

A continuación, se presentan dos diagramas con los que se definen los pasos para medir o calcular la permeabilidad (desde la perspectiva de la Industria Petrolera), ya sea en laboratorio o campo. Para los fines de este trabajo de tesis se pone especial énfasis en los datos obtenidos en campo derivados de pruebas de presión y de registros de pozo.

Además, se define el proceso por el que pasan los núcleos desde el corte y recuperación (Figura 1.7 y Figura 1.8).

Determinación de la Permeabilidad de la Roca a través de Metodologías Convencionales en la Industria Petrolera.





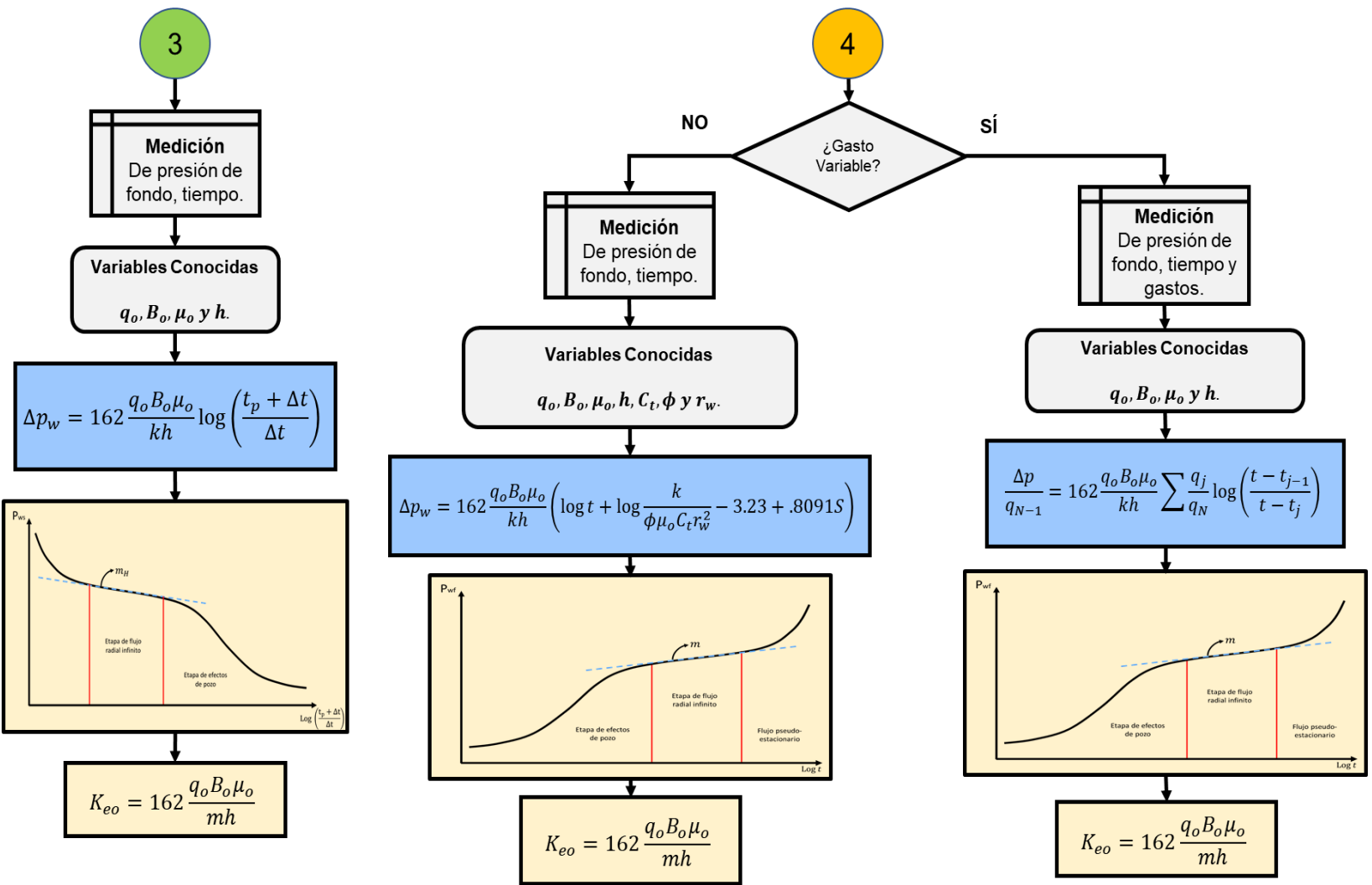


Figura 1.7. Diagrama de Razonamiento de los Procesos Convencionales para la Determinación de la Permeabilidad.

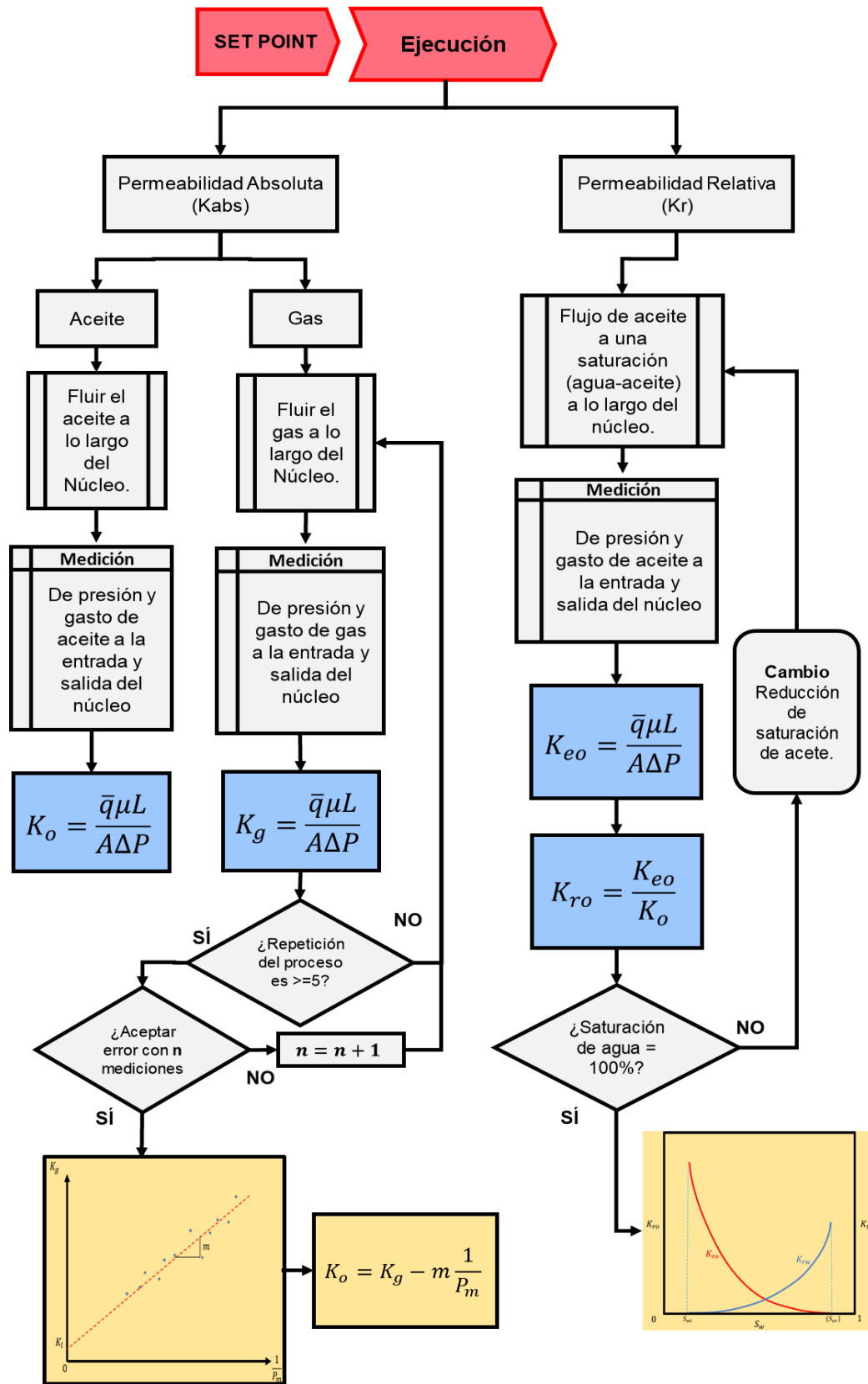


Figura 1.8. Diagrama de Razonamiento de los Procedimientos en Laboratorio.

1.3.1 Pruebas Transitorias de Presión.

Una prueba transitoria de presión ya sea incremento, decremento o sus variaciones (vea **Figura 1.9**), incluye la generación y medición de la variación de presión con el tiempo dentro del pozo y, subsecuentemente, la estimación de las propiedades de la roca, del fluido e incluso del mismo pozo.

Típicamente en este tipo de pruebas se busca determinar la permeabilidad a escala de yacimiento por lo que se realizan en distintos momentos a lo largo de la vida de un proyecto petrolero: durante el proceso de exploración y desarrollo del campo con pruebas DST (por sus siglas en inglés *drill stem test*) o durante el periodo de explotación.

Las pruebas de incremento de presión (**Figura 1.9a**) requieren del cierre un pozo productor. Las técnicas más simples de análisis piden que el pozo produzca a un gasto constante (q_{prod}), ya sea desde el inicio de su vida productiva o durante el tiempo suficiente para establecer una presión estabilizada (conocida como p_{ss}) antes de dicho cierre (Earlougher, 1977). El t_p representa el tiempo de producción y Δt es el tiempo medido desde el cierre del pozo.

En las pruebas de decremento de presión (**Figura 1.9b**) se hacen mediciones de la presión de fondo del pozo durante un periodo de flujo a un gasto de producción constante (q_{prueba}). Usualmente el pozo es cerrado con anterioridad a este periodo de flujo por un cierto tiempo para permitir que la presión se estabilice a lo largo de toda la formación, una vez logrado esto, el pozo es abierto al tiempo t_0 después del cual se comienza con las mediciones. Para aquellas pruebas donde no es posible mantener un gasto constante o bien, donde el pozo no se cerró el tiempo suficiente, existen métodos de análisis que consideran un gasto variable a lo largo de toda la prueba (Matthews & Russel, 1967).

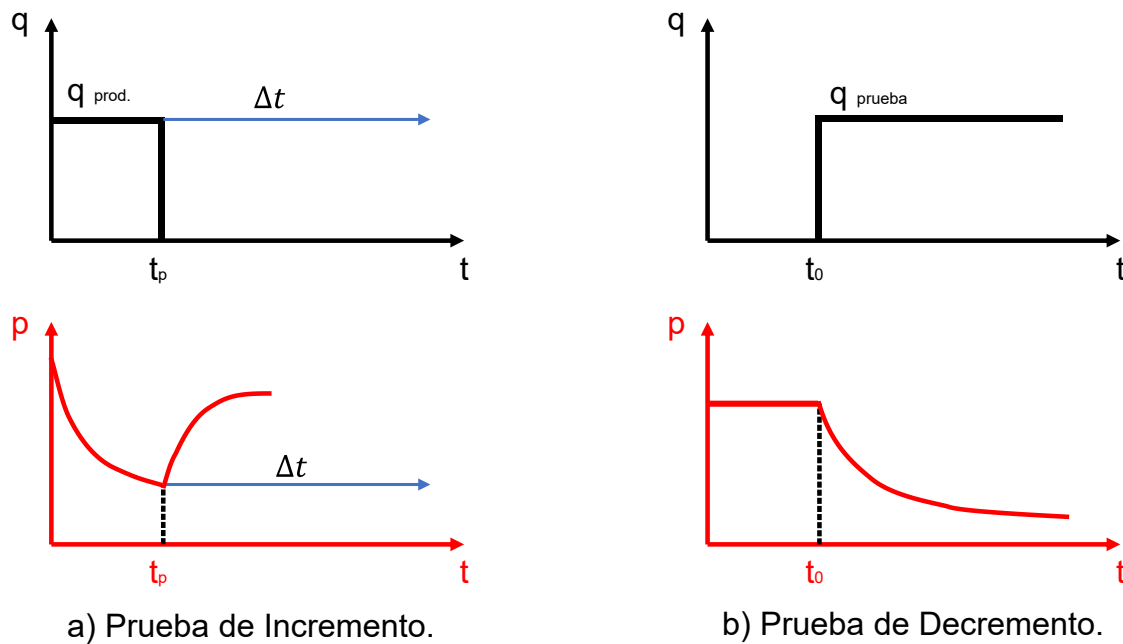


Figura 1.9. Diagramas de comportamiento de la presión y el gasto en una prueba de incremento (a.) y decremento (b.). Modificado de (Lee, 1982).

Análisis de Pruebas de Presión.

Para una prueba de incremento de presión se hace uso del gráfico semilogarítmico especializado de Horner (Horner, 1951), asumiendo que es un sistema que se encuentra en una geometría de flujo radial (**Figura 1.10**). En este gráfico se ubica la sección de los datos de presión que forman una línea recta (línea punteada azul) lo que es indicativo de que éstos no se encuentran afectados por las fronteras ni los efectos de almacenamiento del pozo, una vez ubicada la sección recta de la curva se determina la pendiente (m_H) de la misma. La ecuación 1.6 sirve como base para la determinación de la permeabilidad efectiva al aceite (o al fluido extraído) (k_{eo}):

$$P_{ws} = P_i - 162.6 \frac{qB\mu}{k_{eo}h} \log\left(\frac{t_p + \Delta t}{\Delta t}\right), \quad \dots \quad (1.6)$$

donde

P_{ws} = Presión de fondo estática [psi].

P_i = Presión inicial [psi].

$q =$ Gasto antes del cierre [bpd].

$B =$ Factor de volumen [1].

$\mu =$ Viscosidad [cP].

$k_{eo} =$ Permeabilidad efectiva al aceite [md].

$h =$ Altura o grosor de la formación que contiene hidrocarburos [ft].

$t_p =$ Tiempo de producción antes del cierre.

$\Delta t =$ Tiempo al que se realizó la medición de presión después del cierre [horas].

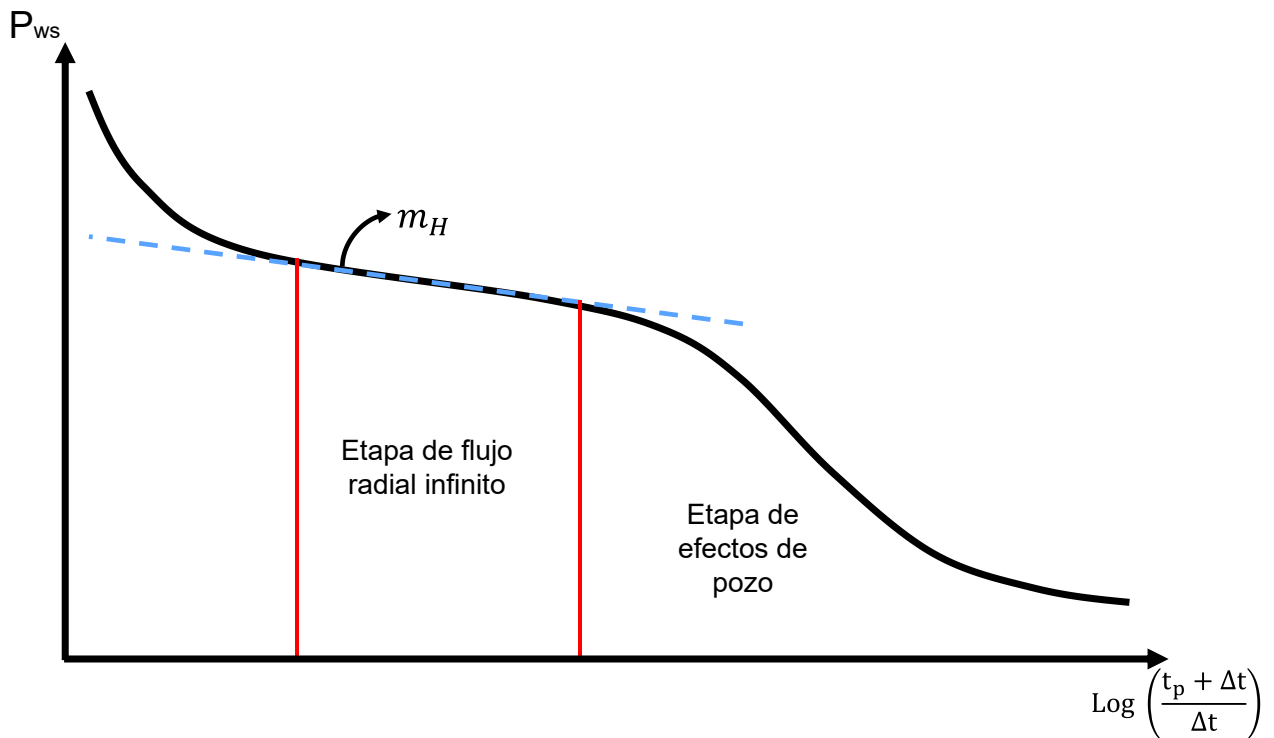


Figura 1.10. Gráfico Especializado de Horner.

De la ecuación 1.6 se deduce que la pendiente se encuentra dada de la siguiente forma:

$$m_H = -162.6 \frac{qB\mu}{k_{eo}h} \dots\dots\dots (1.7)$$

Realizando el despeje correspondiente de la permeabilidad efectiva al aceite (k_{eo}):

$$k_{eo} = -162.6 \frac{qB\mu}{m_H h} \dots\dots\dots (1.8)$$

Para las pruebas de decremento de presión se hace uso de la solución línea fuente para la ecuación de difusividad en coordenadas radiales. Para calcular el valor de permeabilidad a partir de los datos de una de estas pruebas se deben tener presentes las siguientes consideraciones: existe flujo radial infinito, el tamaño del yacimiento es mucho mayor al radio del pozo que éste se considera despreciable (muy cercano a cero) y la presión medida en una sección del yacimiento muy lejana al pozo es igual a la presión inicial del yacimiento (Lee, 1982).

En la **Figura 1.11** se presenta el gráfico para este tipo de pruebas. El proceso de análisis es similar al gráfico de Horner. Se determina la sección de los datos de presión que no se encuentra afectada por efectos tempranos de almacenamiento, daño o la descarga de la tubería ni por aquellos efectos tardíos correspondientes al periodo donde la onda de presión provocada por la prueba conoce las fronteras del yacimiento, la cual tiene un comportamiento identificable debido a la caída de presión presente una vez se llega a esta condición de flujo. De lo anterior, que se busca la sección de los datos que formen una línea recta, de esta es posible obtener la pendiente (m), esta línea recta identifica los datos de presión que se encuentran en un estado transitorio en un flujo radial infinito para que el resultado obtenido de este análisis se considere representativo y correcto.

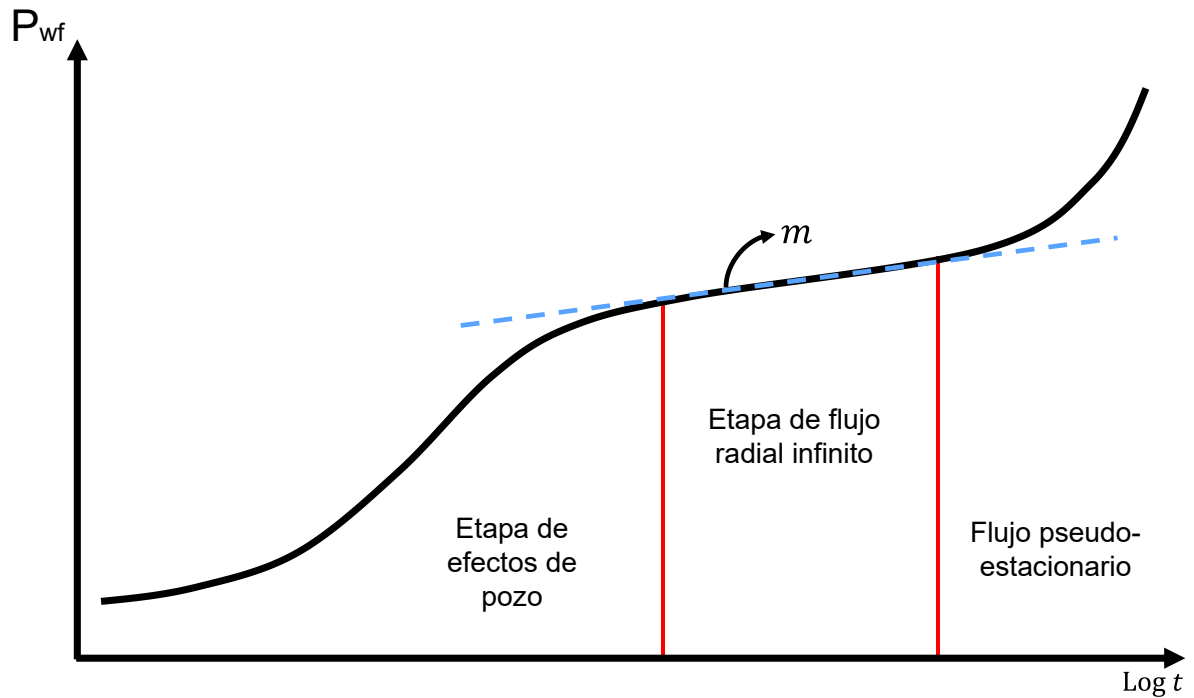


Figura 1.11. Gráfico Especializado para Flujo radial Infinito.

La ecuación usada para este caso es:

$$P_{wf} = P_i - \frac{162.6qB\mu}{K_{eo}h} \left[\log t + \log \left(\frac{k}{\phi\mu C_t r w^2} \right) - 3.227 + .87S \right]. \quad \dots\dots\dots (1.9)$$

Identificando la pendiente para esta ecuación se tiene que:

$$m = 162.6 \frac{qB\mu}{k_{eo}h}. \quad \dots\dots\dots (1.10)$$

Finalmente despejando se llega a:

$$k_{eo} = 162.6 \frac{qB\mu}{mh}. \quad \dots\dots\dots (1.11)$$

1.3.2 Modelos Semi-empíricos

Debido a que el tamaño y distribución de los granos de roca en un medio poroso están relacionados al área que ocupan y esto afecta a la permeabilidad de dicho medio, esto ha sido usado para crear ecuaciones empíricas para la predicción de la permeabilidad (Shang et al, 2003).

Estos modelos pueden dividirse en aquellos considerados convencionales y los relacionados a los estudios de resonancia magnética nuclear (RMN). Debido a los alcances del trabajo, a continuación, se detallan los modelos semi-empíricos convencionales más sólidos y utilizados (**Tabla 1.1**).

Tabla 1.1. Métodos Semi-empíricos para la Determinación de la Permeabilidad (Mohaghegh, 1997).

Investigador	Año	Ecuación
Tixier	1949	$\left(\frac{k}{20}\right)^{1/2} = \frac{2.3}{R_o(\rho_w - \rho_o)} \frac{\Delta R}{\Delta D}$
Timur	1968	$k = 0.136 \frac{\phi^{4.4}}{S_{wi}^2}$
Coates & Dumanoir	1974	$k^{1/2} = \frac{C}{W^4} \frac{\phi^{2W}}{R_w/R_i}$
Coates	1981	$k^{1/2} = 100 \frac{\phi^{2(1-S_{wi})}}{S_{wi}}$

Todos estos métodos a excepción del de Coates & Dumanoir asumen ciertos valores para el factor de cementación y/o del exponente de saturación y son aplicables a formaciones de arenas limpias donde existen condiciones de saturación de agua residual (Mohaghegh, 1997).

Tixier en el año de 1949 (Tixier, 1949) usando relaciones empíricas entre la resistividad, la saturación de agua y la presión capilar estableció un método para calcular la permeabilidad a partir de gradientes de resistividad. Este método asume que el exponente de saturación n , es igual a 2.0. El modelo está limitado por la escasez de registros que exhiben el contacto agua-aceite y por la necesidad de estimar la densidad del fluido encontrado en el yacimiento. Además, la permeabilidad calculada es un promedio de la zona donde se obtuvo al gradiente de resistividad.

Timur (Timur, 1968) propuso una ecuación generalizada para areniscas. Se basó en 155 mediciones de núcleos de esta litología de distintos campos. Esta ecuación (mostrada en la **Tabla 1.1**) relaciona a la permeabilidad con la porosidad y la saturación de agua irreductible en el medio poroso. En la práctica, puede obtenerse una gran diferencia entre los valores medidos y estimados por esta ecuación, por lo que se recomienda ajustar la potencia 4.4 mostrada en la ecuación (Shang et al, 2003).

Coates y Dumanoir en 1974 (Coates & Dumanoir, 1974) propusieron una técnica empírica mejorada para el cálculo de la permeabilidad. Con la ayuda de datos de registros petrofísicos, formularon un exponente en común w , tanto para el exponente de saturación n , como para el exponente de cementación m . Donde w es calculado como sigue:

$$w^2 = (3.75 - \phi) + \frac{1}{2} [\log_{10} \left(\frac{R_w}{R_{ti}} \right) + 2.20]^2, \quad \dots\dots\dots (1.12)$$

donde

$w =$ exponente w .

$\phi =$ porosidad de la roca.

$R_w =$ Resistividad de la Salmuera.

$R_{ti} =$ Resistividad de la formación a la saturación de agua irreductible.

Estas ecuaciones son válidas para formaciones limpias que tengan saturación de aceite con una densidad igual a 0.8. Ellos también presentaron una metodología para probar si en la formación existía saturación de agua irreductible, sin embargo, notaron que para un medio heterogéneo dicha metodología podría fallar. Este método es el único, hasta 1974, que satisface la condición de cero-permeabilidad, cero-porosidad cuando existe una saturación de agua irreductible del 100%.

Coates en 1981 propuso una expresión para la determinación de la permeabilidad que igualmente satisface la condición de cero-permeabilidad, cero-porosidad. Para su uso óptimo la formación tiene que estar a una saturación de agua irreductible del 100% (Balan, Mohaghegh, & Ameri, 1995).

Análisis en Laboratorio

Un programa de obtención de núcleos es muy similar a muchos proyectos de ingeniería pues tiene la premisa de que la inversión tendrá un beneficio. Inicia con la fase de exploración de fuentes alternas de información: pruebas de pozo, registros petrofísicos, muestras anteriores de núcleo y muestras de pared (Exploration and Production Department API, 1998).

En una planeación correcta del programa deben enlistarse completamente los objetivos (generales y específicos) como: identificar litologías, ambientes de depósito, tipo de poro, información sobre la permeabilidad y su relación con la porosidad, presiones capilares, así como información para refinar cálculos que necesiten de registros de pozo.

Lo siguiente es seleccionar el fluido de perforación en función de las necesidades del proyecto de nucleación y de las características de la formación. Seleccionado el fluido se procede a realizar el corte en la roca mediante los llamados *barriles de corte para núcleo*. Estos barriles están diseñados para recuperar el núcleo cortado en lo profundo de la formación manteniendo las condiciones originales los fluidos del yacimiento. Estos barriles de corte pueden clasificarse en convencionales y especiales (Pillado Torres, 2016).

Barriles de Corte Convencionales.

Este tipo de herramientas están disponibles para cortes de diámetros que van desde 1.75 hasta 5.25 pulgadas (44.5 a 133.4 milímetros), teniendo una longitud que puede ir de 1.5 pies (.46 metros) para aplicaciones en pozos horizontales de radio pequeño hasta 400 pies (121.9 metros) para formaciones de mayor grosor que además estén bien consolidadas y sean mayormente homogéneas. El diámetro y longitud de los núcleos se encuentra en función del tamaño y ángulo de inclinación del pozo, así como los esfuerzos sobre la roca y la litología de la formación. En la **Tabla 1.2**, se resumen los sistemas convencionales para el corte y obtención de núcleos, el barril interno, la longitud que puede ser muestreada y sus características especiales.

Tabla 1.2. Sistemas Convencionales para corte de Núcleos. Modificado de (*Exploration and Production Department API, 1998*).

Barril Interno	Longitud	Características Especiales
Acero dulce.	30 - 120 ft (9.14 - 36.58m).	Para altas temperaturas. Sistema de preservación.
Acero dulce.	1.5 ft (.46m).	Para radios de núcleo pequeños.
Acero de alto esfuerzo.	120 - 400 ft (36.38 – 121.9 m).	Barril fuerte, incluye estabilizador.
Fibra de vidrio.	30 – 90 ft (9.14 – 27.43 m).	Para formaciones consolidadas y no consolidadas.
Aluminio.	30 – 90 ft (9.14 – 27.43 m).	Temperatura máx. de 350 F (176.6°C)
Acero con liner de plástico.	30 ft (9.14 m).	Reducción del diámetro del núcleo a ½, Temperatura máx. de 180 F (82.2°C).
Acero con liner de fibra de vidrio.	30 ft (9.14 m).	Reducción del diámetro del núcleo a ½, Temperatura máx. de 250 F (121°C).
Acero con liner de aluminio.	30 ft (9.14 m).	Reducción del diámetro del núcleo a ½, Temperatura máx. de 350 F (176.7°C).

La **Figura 1.12** presenta los componentes comunes de este tipo de barriles de corte.

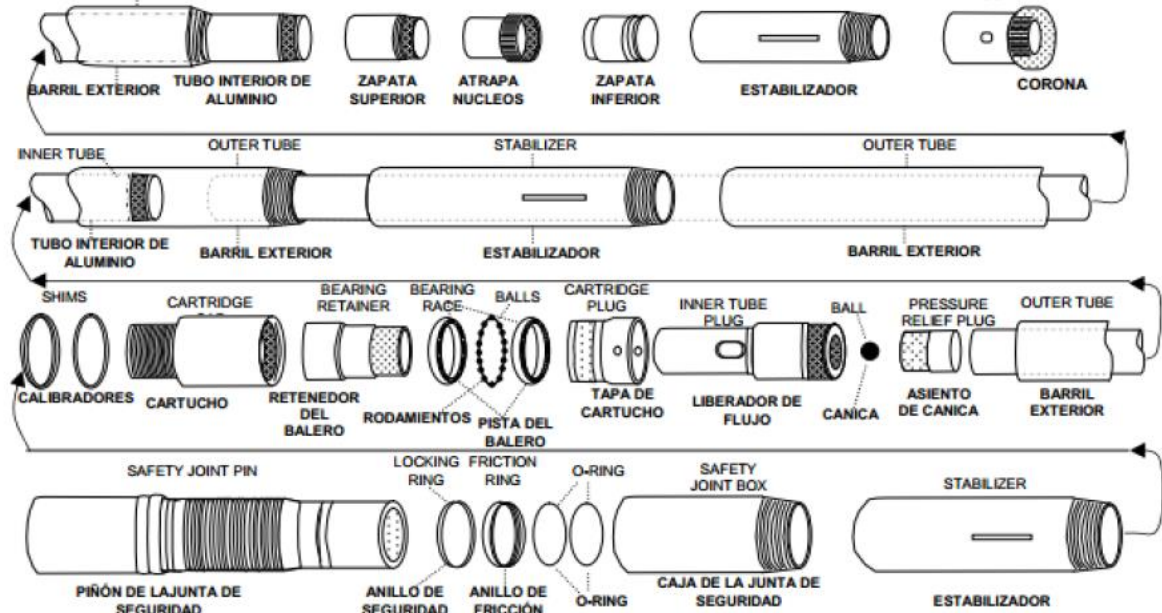


Figura 1.12. Partes Internas de un Barril Nucleador Convencional. Tomada de (Pillado Torres, 2016).

Barriles de Corte Especiales.

Estas herramientas han evolucionado para cubrir necesidades específicas en la obtención de núcleos. Barriles presurizados y aquellos recubiertos de esponja surgieron de la necesidad de mejores datos de la saturación original de aceite del núcleo tomado, mientras que los sistemas de manga de goma y sellado hermético fueron diseñados para mejorar la calidad de los núcleos obtenidos de formaciones no consolidadas y con una alta heterogeneidad (**Tabla 1.3**).

Tabla 1.3. Sistemas de Corte Especiales para Núcleos. Modificado de (*Exploration and Production Department API, 1998*).

Barril de Corte	Dimensión Máxima de Núcleo.	Aplicación Especial.
Barril presurizado.	3.75 in por 10 ft a 5000 psi. 2.5 in por 20 ft a 10000 psi	Análisis a las condiciones presurizadas, saturación de fluidos, volumen y composición de gas.
Liner-esponja.	3.5 in por 30 ft.	Saturación de fluidos.
Cierre total.	4 in por 60 ft	Recuperación en formaciones no-consolidadas.
Con mangas de hule.	3 in por 20 ft	Recuperación en formaciones no-consolidadas, fracturadas, o compuestas por conglomerados.
Recuperable con línea de acero.	2.75 in por 30 ft	Extracción de la muestra sin la sarta de perforación.
Pared de pozo por percusión.	1 in por 1.75 in.	Muestras obtenidas después de las actividades de perforación y toma de registros.
Pared de pozo por rotación.	.94 in por 1.75 in.	Muestras obtenidas después de las actividades de perforación y toma de registros.
Nucleado de pared.	2.5 in por 10 ft.	Núcleo obtenido después de las actividades de perforación y toma de registros.

La **Figura 1.13** muestra un esquema de los componentes comunes de un barril de corte no-convencional a presión.

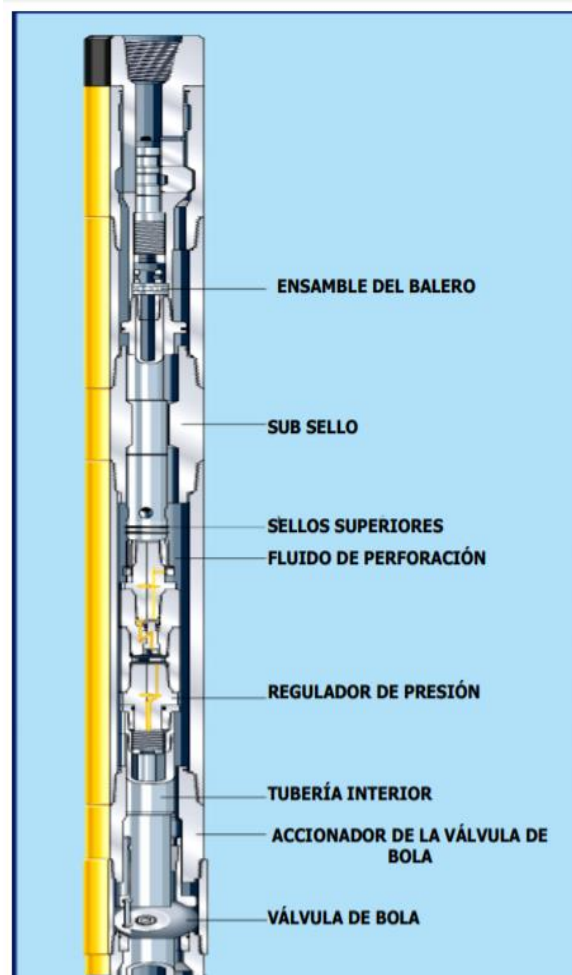


Figura 1.13. Barril no-Convencional de corte. Tomado de (Pillado Torres, 2016).

Manejo y Etiquetado.

Una vez cortado y asegurado el núcleo o la muestra, se procede a llevarla a superficie para su registro y etiquetado, los cuales deben seguir procesos de preservación y manejo que asegure la representatividad de las condiciones del yacimiento. El manejo del núcleo en superficie tiene los siguientes objetivos.

- a) Obtener material rocoso representativo de la formación.
- b) Minimizar la alteración física del material rocoso durante la manipulación y almacenamiento de éste.

Entre los mayores problemas que puede acarrear el cumplir los objetivos anteriormente mencionados se encuentra: la selección de un material no-reactivo y un método para prevenir la pérdida de fluidos o la absorción de contaminantes y la apropiada aplicación de los métodos de manejo y preservación dependiendo del tipo de roca, el grado de consolidación y el tipo de fluido que ésta contiene.

Diferentes tipos de roca podrían requerir precauciones adicionales para lograr la obtención de información representativa. Todos los materiales nucleados deben ser preservados lo antes posible en la zona de perforación una vez se han recuperado para minimizar la exposición a las condiciones atmosféricas.

En caso de ser posible y si se cuenta con el espacio disponible, el núcleo debe ser empaquetado en el piso de las instalaciones de perforación. Se debe tener cuidado de mantener la orientación, y de preservar de forma correcta la secuencia de los núcleos. El punto clave es que el núcleo debe estar registrado y marcado de tal forma que el intervalo completo muestreado pueda ser reensamblado en un tiempo futuro. El núcleo debe ser protegido contra temperaturas extremas, humedad y deshidratación (Exploration and Production Department API, 1998).

Preservación.

La preservación de un núcleo se realiza como un intento para conservarlo, antes del análisis, en la misma condición a la que se encontraba previo a su remoción del barril de corte. En el proceso de cortar el núcleo, resguardarlo y llevarlo a superficie, el contenido de fluido en la roca se altera invariablemente por cambios en la presión, temperatura, etc. Prácticas incorrectas o sin cuidado durante el manejo y la preservación causan una mayor alteración del núcleo y sus fluidos haciéndolos aún menos representativos de la formación. La preservación y empaquetamiento de los núcleos varía dependiendo de las pruebas a realizar, el tiempo antes de dichas pruebas y el potencial de realizar análisis en el sitio de la perforación. Si las muestras de núcleo serán usadas para determinar la saturación de fluidos o para un análisis especial, es necesario que estos sean preservados para su transporte al laboratorio. La evaporación y migración de los fluidos, así como la oxidación dentro de la muestra debe evitarse para obtener un análisis confiable del núcleo. Núcleos provenientes de formaciones consolidadas bien pueden

durar lo suficiente para no requerir procedimientos especiales de manejo, sin embargo, se debe tener especial cuidado con ejemplares no-consolidados, fracturados, etc (Parrado, 2016).

No se recomienda el uso de recipientes de vidrio, plásticos deformables, cartón, contenedores no rígidos ni de envases con aire presurizado con propósito de preservación.

No existe un mejor método de preservación. La experiencia puede ayudar a determinar el método más satisfactorio según el tipo de roca. El método para elegir dependerá de la composición, grado de consolidación y características distintivas de la roca. Las técnicas requeridas para preservar núcleos para pruebas dependerán del tiempo de transporte, almacenaje y la naturaleza de la prueba a realizar. Los métodos preferidos para preservar núcleos que serán analizados en laboratorio incluyen uno o más de los listados a continuación (Pillado Torres, 2016):

- a) Estabilización mecánica.
- b) Preservación mediante ambiente controlado utilizando enfriamiento, humedad regulada o hasta congelamiento de ser necesario.
- c) Láminas plásticas selladas con calor.
- d) Bolsas plásticas.
- e) Inmersión y revestimiento de resina.
- f) Sellados en barriles internos desechables, liners y tubos.
- g) Recipientes anaeróbicos (al vacío).

La **Figura 1.14**, muestra un esquema de una caja utilizada para la preservación y transporte de núcleos los cuales, además, se encuentran empacados con papel vinitel (láminas plásticas).

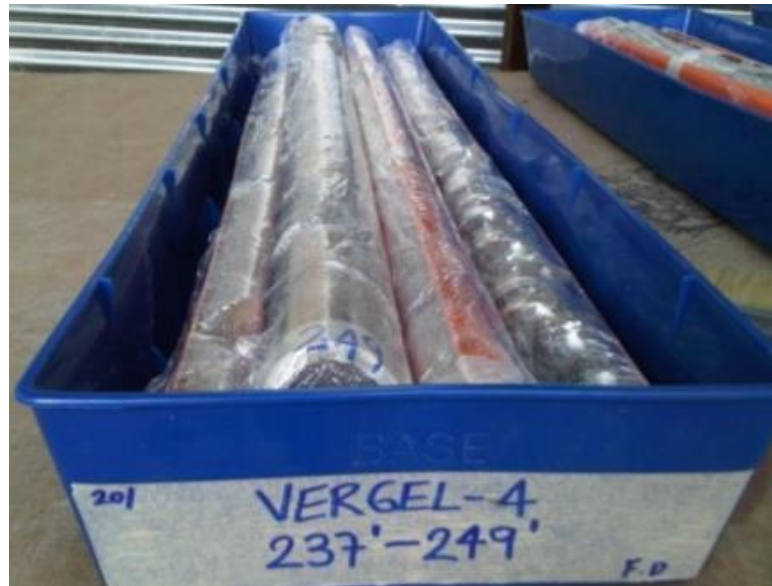


Figura 1.14. Empacado de núcleos con papel vinitel. Tomado de (Parrado, 2016).

Transporte al laboratorio.

El método de transporte debe proveer protección contra el daño por cambios ambientales, vibración mecánica y mal manejo. Otros factores importantes que considerar cuando se elige el modo de transporte incluyen, a) distancia del laboratorio y qué tan remoto se encuentra el sitio donde se realizó la perforación, b) terreno y condiciones terrestres o marítimas, c) condiciones climatológicas, d) tipo de empaquetamiento y preservación y e) costo. En cualquier caso, se deben tomar precauciones para estabilizar el núcleo transportado.

Preparación y descripción del núcleo.

Cuando el núcleo llega al laboratorio, pequeñas muestras conocidas como *plugs* son cortadas cada 20 o 30 cm a lo largo del intervalo productor. Todas estas muestras son analizadas respecto a la porosidad, permeabilidad, saturación y litología.

El propósito de preparar y describir el núcleo, además de lo antes mencionado, es el de reconocer el ambiente deposicional, así como características estructurales y diagenéticas de éste. Estas descripciones cualitativas y cuantitativas del núcleo proveen la base para posteriores análisis como registros gamma y gamma espectral de núcleo, imágenes de núcleo, fluoroscopia, radiografía-x, tomografía computarizada y resonancia magnética

nuclear. Los resultados obtenidos son usados para interpretar y evaluar el yacimiento (Torsaeter & Abtahi, 2003).

La **Tabla 1.4**, muestra el procedimiento de un análisis de núcleo de rutina con algunas mediciones consideradas suplementarias.

Tabla 1.4. Análisis Rutinario y Mediciones Adicionales. Modificado de (Torsaeter & Abtahi, 2003).

Propiedad	Aplicación
Análisis Rutinario de Núcleos	
Porosidad	Capacidad de Almacenamiento del medio.
Permeabilidad	Capacidad de flujo.
Saturaciones	Definición de los hidrocarburos móviles (zonas productivas y contactos), tipo de hidrocarburos.
Litología	Tipo de roca y características.
Análisis Adicionales	
Permeabilidad Vertical	Efectos drene gravitacional, conificación etc.
Registro Superficial	Identificación de secciones de núcleo perdidas, relacionar
Gamma	núcleos y registros.
Matriz de Densidades	Calibración del registro de densidad.
Análisis de Aceite y Agua	Densidades, viscosidades, tensión interfacial, composición etc.

El análisis especial de núcleos incluye una gran cantidad de mediciones con el objetivo de obtener información detallada sobre el comportamiento del flujo multifásico. De estos análisis se obtiene información de la distribución de los fluidos en el yacimiento, saturación residual de aceite y características del flujo multifásico (permeabilidades relativas). Ocasionalmente se incluye la medición de propiedades eléctricas y acústicas en este tipo de estudios.

En la **Tabla 1.5**, se incluye una variedad de pruebas especiales de núcleo, así como la información obtenida de estas.

Tabla 1.5. Análisis Especial de Núcleos. Modificado de (Torsaeter & Abtahi, 2003).

Pruebas / estudios	Información / Propiedades
Pruebas Estáticas	
Estudios de Compresibilidad	Permeabilidad y porosidad contra presión
Estudios Petrográficos	Identificación mineralógica, diagénesis, identificación de arcillas, tamaño de grano, presión capilar, geometría de poro etc.
Mojabilidad	Ángulo de contacto e índice de mojabilidad.
Capilaridad	Presión capilar y saturación.
Pruebas Acústicas	
Pruebas Eléctricas	
Pruebas Dinámicas	
Estudios de Flujo	Permeabilidad relativa y puntos finales de saturación.
Estudios de Flujo-EOR	Inyectividad y saturaciones residuales

Antes de la mayoría de las mediciones de porosidad y permeabilidad en laboratorio, los fluidos originales deben ser completamente removidos de la muestra de núcleo. Generalmente, esto se logra a través de procedimientos como el *flushing* (desplazamiento) que consiste en inyectar solventes para extraer hidrocarburos, agua y salmuera (Exploration and Production Department API, 1998).

El solvente puede ser inyectado a la muestra por presión directa. La extracción de hidrocarburo y salmuera es realizada mediante la inyección de uno o varios solventes dentro de la muestra de núcleo bajo presión y a temperatura ambiente. La presión usada dependerá de la permeabilidad de la muestra y varía en un rango de entre 10 hasta 1000 psi. El volumen de solvente necesario para remover completamente los hidrocarburos de la muestra dependerá de los hidrocarburos saturándola y del tipo de solvente usado. Se considera que el núcleo está limpio cuando el flujo de solvente que sale de éste se encuentra sin rastro de hidrocarburos. En algunos casos más de un solvente será necesario para remover los crudos pesados y de tipo asfáltico (Andersen & Duncan, 2013).

Finalmente, el secado de un núcleo o la muestra de éste puede darse mediante cualquiera de los siguientes métodos.

- Horno convencional al vacío.
- Horno de humedad con 40% de humedad relativa.

1.3.3 Medición de la Permeabilidad en Laboratorio.

La medición de la permeabilidad de un medio poroso, o de un estrato, es la medida de la conductividad del fluido de ese material en particular, de aquí que esta medición del flujo a través de una muestra en una dirección dada sirva para el cálculo de la permeabilidad en dicha dirección (Monicard, 1980). La permeabilidad de un medio homogéneo e isotrópico es la misma en todo el medio y hacia cualquier dirección, sin embargo, las rocas reales no son perfectamente homogéneas ni anisotrópicas.

El procedimiento general para medir la permeabilidad consiste en hacer pasar un fluido de viscosidad conocida a través de una muestra de núcleo de dimensiones igualmente conocidas, para después medir el gasto y la caída de presión. Distintas técnicas son utilizadas para medir la permeabilidad de las muestras de núcleos, estas dependen de:

- Las dimensiones y forma de la muestra.
- Estado de consolidación.
- Tipo de fluido usado.
- Rangos de confinamiento y presión del fluido aplicada, y
- rangos de permeabilidad del núcleo.

Dos tipos de instrumentos son comúnmente usados en el laboratorio:

- Permeámetro de carga variable, tipo IFP.
- Permeámetro de carga constante, tipo de Laboratorio de Núcleo.

El permeámetro de carga constante mostrado esquemáticamente en la **Figura 1.15**, es usado regularmente cuando se busca medir permeabilidades de un núcleo completo o de un *plug* obtenido de éste, además, es viable realizar experimentos tanto de una fase como multifásicos. Finalmente, es posible realizar mediciones utilizando fluido

compresible (gas) o algún líquido, es importante resaltar que estas pueden realizarse a condiciones de presión y temperatura del yacimiento.

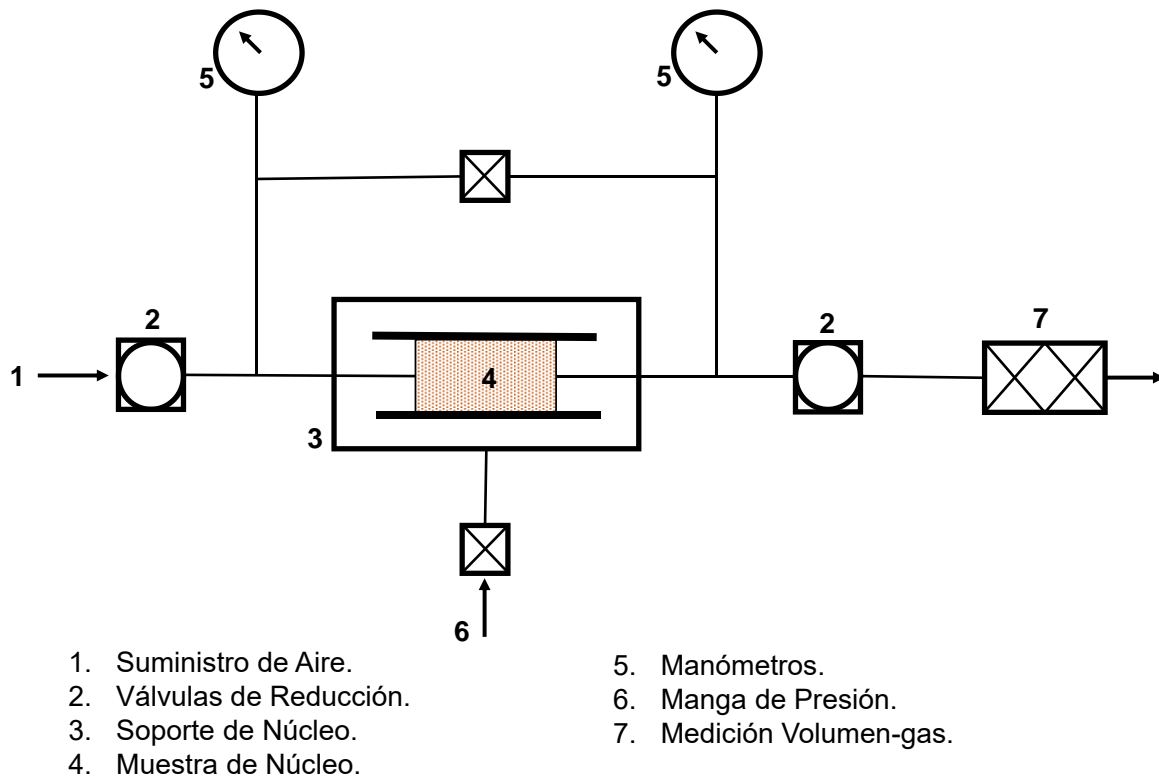


Figura 1.15. Permeámetro de Carga Constante. Modificado de (Torsaeter & Abtahi, 2003).

Como se mencionó, estas pruebas deben realizarse en muestras que hayan sido limpiadas y secadas con anterioridad, esto es porque:

- El estado estacionario en el flujo del fluido a través del núcleo es alcanzado rápidamente.
- El aire seco no altera los minerales en la roca, y
- se llega fácilmente a una saturación del 100% del fluido.

Medición de la Permeabilidad Absoluta

El objetivo de este experimento es el de medir la permeabilidad absoluta (k o k_a) de un fluido usando un método basado en la teoría de Darcy.

Cumplidos los requerimientos anteriormente mencionados correspondientes al procedimiento general de medición, se hace fluir el fluido seleccionado (agua o aceite) a través de la muestra a un gasto constante, es necesario que el flujo alcance un estado estacionario para poder realizar las mediciones de caída de presión en la longitud seleccionada de la muestra y el gasto a la salida, posteriormente se utiliza de la Ley de Darcy con la ecuación mostrada a continuación para calcular la permeabilidad absoluta de la muestra.

$$k = \frac{q_f \mu L}{A \Delta P} \quad \text{si } S = 100\%, \quad \dots \dots \dots (1.13)$$

donde

k = Permeabilidad absoluta de la muestra [L^2].

q_f = Gasto a la salida de la muestra [L^3/t].

μ = Viscosidad del fluido utilizado [m/Lt].

L = Longitud de la muestra [L].

A = Área transversal de la muestra [L^2].

ΔP = Caída de presión medida a lo largo de la longitud L de la muestra [F/L^2].

El procedimiento previamente descrito puede llevarse a cabo utilizando gas como fluido saturante. En análisis de núcleos rutinarios puede usarse aire, pero los gases inertes como el nitrógeno y el helio son más comunes.

El flujo de gas en los poros es mayor que el flujo de líquido debido a que el líquido experimenta una mayor resistencia, o arrastre, en las paredes de poro que aquel que experimentan los gases. Este resbalamiento es un efecto que puede ser corregido incrementando la presión media del gas en la muestra, lo que comprime el gas y aumenta el arrastre en las paredes de los poros. Por lo tanto, es necesario realizar la corrección de Klinkenberg, que es una extrapolación al infinito de las mediciones de gas, en donde, a presiones muy altas, el gas asume un comportamiento como el del líquido (Klinkenberg, 1941).

En consecuencia, es necesario realizar el procedimiento de fluir el gas a lo largo de la muestra haciendo las mediciones correspondientes un mínimo de cinco veces y tantas como sea posible, según el error que se desee aceptar en la extrapolación. Cabe resaltar que en cada repetición del experimento la presión media en el núcleo deberá ser incrementada. Con estas mediciones se construye el gráfico mostrado en la **Figura 1.16**. En el eje de las x se colocan las mediciones del recíproco de la presión media y en el eje de las y el cálculo de la permeabilidad absoluta utilizando gas resultante de cada una de las mediciones a diferentes presiones medias con la ecuación 1.13.

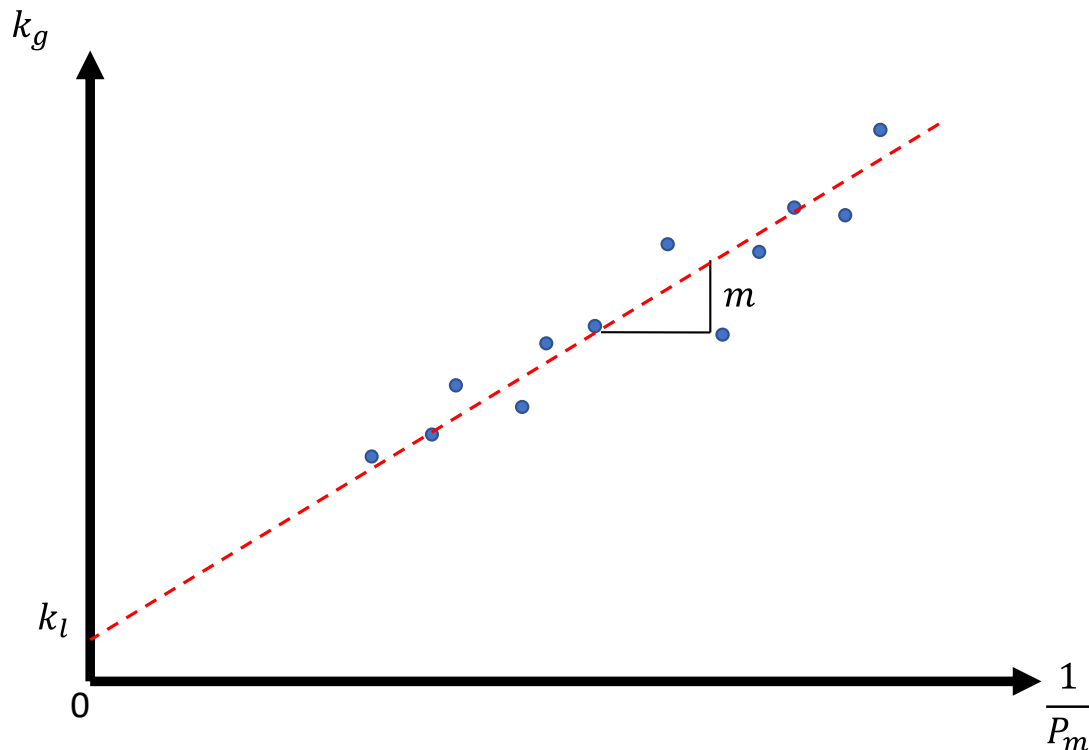


Figura 1.16. Gráfico para la Extrapolación de Klinkenberg.

La ecuación con la que se obtiene la extrapolación de Klinkenberg es la siguiente:

$$k_l = k_g - m \frac{1}{P_m} \dots\dots\dots (1.14)$$

Finalmente, lo que se pretende extrapolar es la ordenada al origen que corresponde con la medición de k_g a una presión que tiende al infinito, determinando así, la permeabilidad absoluta al fluido ($k = k_l$).

Medición de la Permeabilidad Relativa

Como se mencionó anteriormente, la permeabilidad absoluta se refiere a un medio poroso saturado al 100% de un solo fluido. Sin embargo, en los yacimientos petroleros las rocas se encuentran saturadas con dos o más fluidos, como pueden ser agua, aceite y gas (Torsaeter & Abtahi, 2003), es aquí donde crecen en importancia los conceptos de permeabilidad efectiva y permeabilidad relativa.

Para flujos multifásicos, donde fluyen distintos fluidos en un medio poroso parcialmente saturado con cada uno de éstos, cada fase fluye a su propio gasto y compite por los canales de flujo. Su admisión a través del medio poroso se determina por la permeabilidad efectiva. Por otro lado, el flujo fraccional de cada fluido está relacionado con la permeabilidad relativa (Pérez Pacheco, 2011).

En estas pruebas, la muestra puede ser saturada con dos fluidos (agua y aceite o agua y gas) o con tres fluidos (agua, aceite y gas), dados los alcances de este trabajo se pondrá especial interés en el procedimiento de una muestra con una saturación bifásica.

Cuando la muestra se satura con agua y aceite generalmente se comienza teniendo un porcentaje de aceite superior que el del agua, donde dicho porcentaje irá disminuyendo mientras aumenta la saturación del agua. Los fluidos dentro del medio poroso se hacen circular de acuerdo con alguno de los casos (API, 1988):

- a) Un fluido en movimiento y el otro fluido estático.
- b) Los dos fluidos en movimiento.

Cualquiera sea el caso, como en los experimentos anteriores, debe alcanzarse un estado estacionario en el flujo a través de la muestra. Cumplido esto, es necesario medir el gasto a la salida de la muestra (el cual debe permanecer constante durante todo el experimento), la caída de presión en la sección escogida, además de conocer las dimensiones del núcleo y las propiedades de los fluidos manejados. El cálculo de la permeabilidad efectiva a un fluido (k_{ef}) se realiza haciendo uso de la Ley de Darcy como sigue (API, 1988):

$$k_{ef} = \frac{q_f \mu_{fl} L}{A \Delta P}, \quad \dots\dots\dots (1.15)$$

donde

k_{ef} = Permeabilidad efectiva a un fluido [L^2].

q_f = Gasto a la salida de la muestra del fluido [L^3/t].

μ = Viscosidad del fluido [m/Lt].

L = longitud escogida de la muestra [L].

A = Área transversal al flujo [L^2].

ΔP = Caída de presión en la longitud L escogida [F/L^2].

Para el cálculo de la permeabilidad relativa (k_{rf}) y con los valores de permeabilidad efectiva a diferentes saturaciones de los dos fluidos, se usa:

$$k_{rf} = \frac{k_{ef}}{k_a}. \quad \dots\dots\dots (1.16)$$

Generalmente los resultados de permeabilidad relativa en un sistema agua-aceite se presentan de forma gráfica como función de la saturación de agua, la **Figura 1.17**, muestra un ejemplo típico del resultado de un experimento de este tipo.

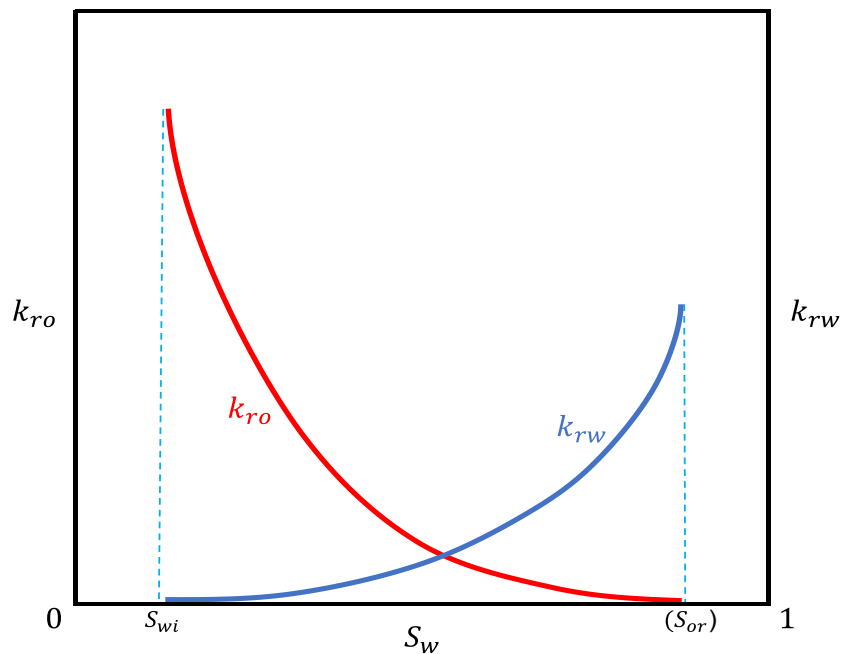


Figura 1.17. Gráfico de Resultados de la Permeabilidad Relativa.

La dirección de las curvas señala la historia de saturaciones mejor conocida como drenaje e imbibición. La curva de *drenaje* aplica a procesos donde la fase mojante decrece en magnitud, mientras que la curva de *imbibición* aplica a aquellos procesos donde la fase no-mojante decrece (McCain, 1990).

Es importante resaltar que las curvas de permeabilidad relativa consisten en tres elementos:

- El punto final de la saturación de fluido.
- El punto final de las permeabilidades, y
- la curva de la función de permeabilidad relativa.

El punto final de las saturaciones determina el rango de saturación móvil y se encuentra directamente relacionada con la cantidad de aceite recuperable. El punto final de las permeabilidades relativas determina la eficiencia de barrido para un proceso de desplazamiento.

La presión de confinamiento a la que las pruebas anteriormente descritas pueden realizarse varía desde la presión atmosférica (14.7 psi) hasta 5000 psi en algunos laboratorios especializados seleccionados (Honarpour et al, 2003).

Capítulo 2: Ciencia de Datos

La Ciencia de Datos, CD, entendida como una disciplina científica, es influenciada por la informática, las ciencias de la computación, las matemáticas, las operaciones de investigación y la estadística, así como por las ciencias aplicadas. Es un campo interdisciplinario acerca de procesos útiles para extraer conocimiento a partir de grandes volúmenes de datos en varios formatos, estructurados o no-estructurados. Puede entenderse como una continuación de algunos campos de análisis de información como la analítica predictiva, el descubrimiento del conocimiento y la minería de datos (**KDD**: Knowledge Discovery and Data mining) (Ranjan, 2016; Grus J., 2015; Piatesky & Frawley, 1991).

Una definición muy simple de la CD es la de Cao (Cao, 2017):

$$CD = (ST + INF + CP + CM + SL + MG)/(DE + TH), \quad \dots\dots\dots (2.1)$$

donde

CD = ciencia de datos.

ST = estadística.

INF = informática.

CP = cómputo.

CM = comunicación.

SL = sociología.

MG = administración.

DE = ambiente de datos.

TH = pensamiento.

En esta fórmula, sociología se refiere a los aspectos sociales y (*DE + TH*) significa que todas las ciencias mencionadas actúan utilizando como base los datos, el ambiente y el pensamiento lógico estructurado: datos → conocimiento → Sabiduría.

Para aplicar CD los pasos a seguir según CRISP-DM (*Cross Industry Standard Process for Data Mining*) (Brown, 2014) son:

- Entendimiento del negocio.
- Entendimiento de los datos.
- Preparación de los datos.
- Modelado.
- Evaluación, y
- despliegue.

Sin embargo, Wehis y Ickstadt (2018) presentan una estructura inspirada en la anterior que se ajusta más a las necesidades y objetivos planteados en este trabajo:

- Adquisición y enriquecimiento de datos.
- Almacenamiento y acceso de datos.
- Exploración de los datos.
- Análisis y modelado de datos.
- Optimización de Algoritmos.
- Selección y validación del modelo.
- Representación y reporte de resultados, y
- despliegue de los resultados.

Usualmente, estos pasos no son realizados una sola vez, sino que se entienden como un proceso iterativo y cíclico. A continuación, se detallan.

Adquisición y enriquecimiento

El diseño de experimentos es esencial para una generación sistemática de datos cuando el efecto de factores ruidosos tiene que ser identificado. Experimentos controlados son fundamentales para robustecer un proceso ingenieril para así lograr productos confiables a pesar de la variación en los procesos.

Exploración de datos

Esta actividad es esencial para el pre-procesamiento de los datos pues de su buena implementación se aprende sobre el contenido de la base. La exploración y visualización de los datos observados, en cierta forma, fue iniciada por John Tukey (Tukey, 1977) y está en continua evolución. Hoy permite entender y transformar lo registrado en un marco sólido como ciencia estadística. La contribución más importante de la estadística es la noción de *distribución*. Esta permite representar la variación en los datos y dirige la selección de modelos y métodos analíticos.

Encontrar la estructura que subyace los datos y realizar predicciones son los pasos más importantes dentro de la Ciencia de Datos, en particular, los métodos estadísticos son esenciales ya que éstos permiten el manejo de diferentes labores analíticas (Weihs & Ickstadt, 2018). A continuación, se enuncian algunos ejemplos importantes de métodos de análisis estadísticos de datos:

- **Prueba de hipótesis:** este es uno de los pilares del análisis estadístico, puesto que los cuestionamientos en problemas manejados por datos pueden a menudo ser trasladados a hipótesis, además, de que estas son el vínculo natural entre la teoría que subyace a los datos y la estadística. Ya que las hipótesis estadísticas se relacionan con las pruebas estadísticas, tanto los cuestionamientos como la teoría pueden ser probados con los datos disponibles (Dmitrienko & Tamhane, 2009).
- **Clasificación:** estos métodos son básicos para encontrar y predecir subgrupos dentro de los datos. En el llamado caso no-supervisado, éstos pueden ser encontrados en un set de datos incluso sin tener conocimiento previo de que existan dichos subgrupos. Convencionalmente a esto se le llama *clustering*. En el caso supervisado, reglas de clasificación deberían ser encontradas a partir de un conjunto de datos.
- **Regresión:** estos métodos son la herramienta principal para encontrar relaciones globales y locales entre los atributos cuando la variable objetivo es medida. Bajo la suposición de que existe normalidad, la regresión lineal es el método más común, mientras que una regresión lineal generalizada es usualmente empleada para otro tipo de distribuciones de la familia exponencial (Fahrmeir L. et al, 2013).

- **Análisis de series de tiempo:** apunta a entender y predecir una estructura temporal (Shunway & Stoffer, 2010). Las series de tiempo son muy comunes en estudios de datos observacionales, y la predicción es el desafío más importante de este tipo de datos. Las áreas donde aplicación más típicamente son las ciencias del comportamiento, economía, así como, las ciencias naturales y las ingenierías (Weihs & Ickstadt, 2018).

Selección y Validación del Modelo

En casos donde más de un modelo es propuesto para predicción, las pruebas estadísticas para comparar estos modelos son de utilidad cuando se tiene en la mira el poder predictivo de cada uno (Vatolkin & Weihs, 2017).

Despliegue y Reporte de los Resultados

La visualización para interpretar las estructuras encontradas y almacenar los modelos en una forma fácil de actualizar son tareas muy importantes para poder comunicar los resultados y salvaguardar el despliegue del análisis de datos. Este despliegue es decisivo para obtener resultados interpretables en la Ciencia de Datos (Berger, 2014).

2.1 Histogramas

Un histograma es un tipo especial de gráfico de barras que representa la distribución de frecuencias de un conjunto de datos, mostrando la acumulación o tendencia, la dispersión y la forma de la distribución de dichos datos (**Figura 2.1**).

Las clases o barras dependerán del modelador por lo que suele ser necesario hacer varios histogramas con diferentes números de clases hasta obtener el que muestre eficientemente las características deseadas.

Es común usarlo para representar variables continuas, aunque también puede aplicarse para variables discretas. Es una herramienta especialmente útil cuando se pretende examinar un gran número de datos dentro de una base, que es necesario organizar para posterior análisis (Riaño R., 2017).

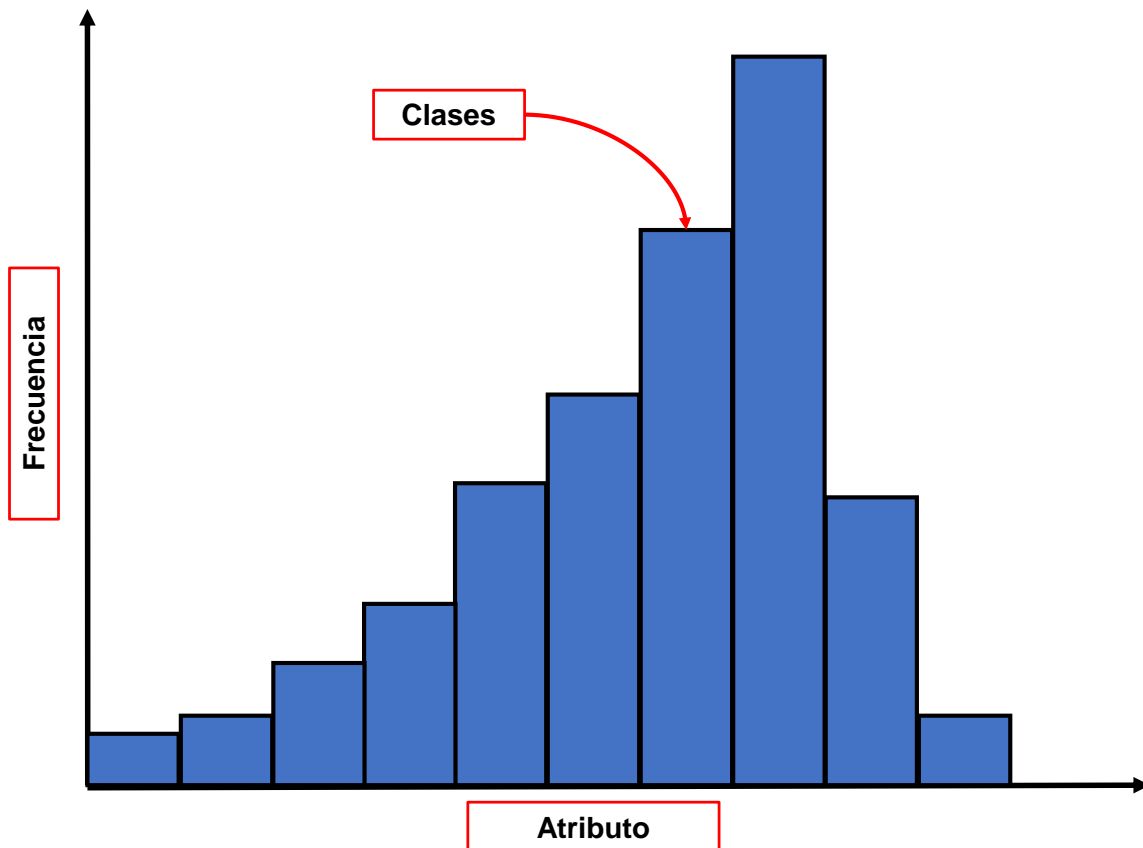


Figura 2.1 Histogramas.

2.2 Diagrama de Cajas y Bigotes

Este tipo de diagrama, también conocido como diagrama de cajas o *boxplot*, es útil para mostrar gráficamente características relativas a la distribución de frecuencias de los datos contenidos en una base de datos (Becerra, ?).

Permite visualizar de un modo intuitivo la mediana, la dispersión de los datos, simetría y aquellos valores atípicos en el rango de valores de un atributo. En caso de observar un sesgo, es posible identificar la dirección de éste.

Un diagrama de cajas y bigotes consiste en un rectángulo atravesado por una línea horizontal que representa la mediana (segundo cuartil). La base de este rectángulo representa el primer cuartil de los datos, mientras que el techo de la caja el tercer cuartil. La distancia entre el primero y tercero de los cuartiles se conoce como rango intercuartil y se manifiesta en el gráfico como la altura de la caja, de esta última salen dos líneas,

también conocidas como bigotes, alineadas con la caja que representan los valores mínimo y máximo que toma la muestra. De este modo se percibe la variabilidad de los datos fuera de los cuartiles uno y tres. En ocasiones se representan también los valores atípicos como puntos aislados en línea con los bigotes (**Figura 2.2**).

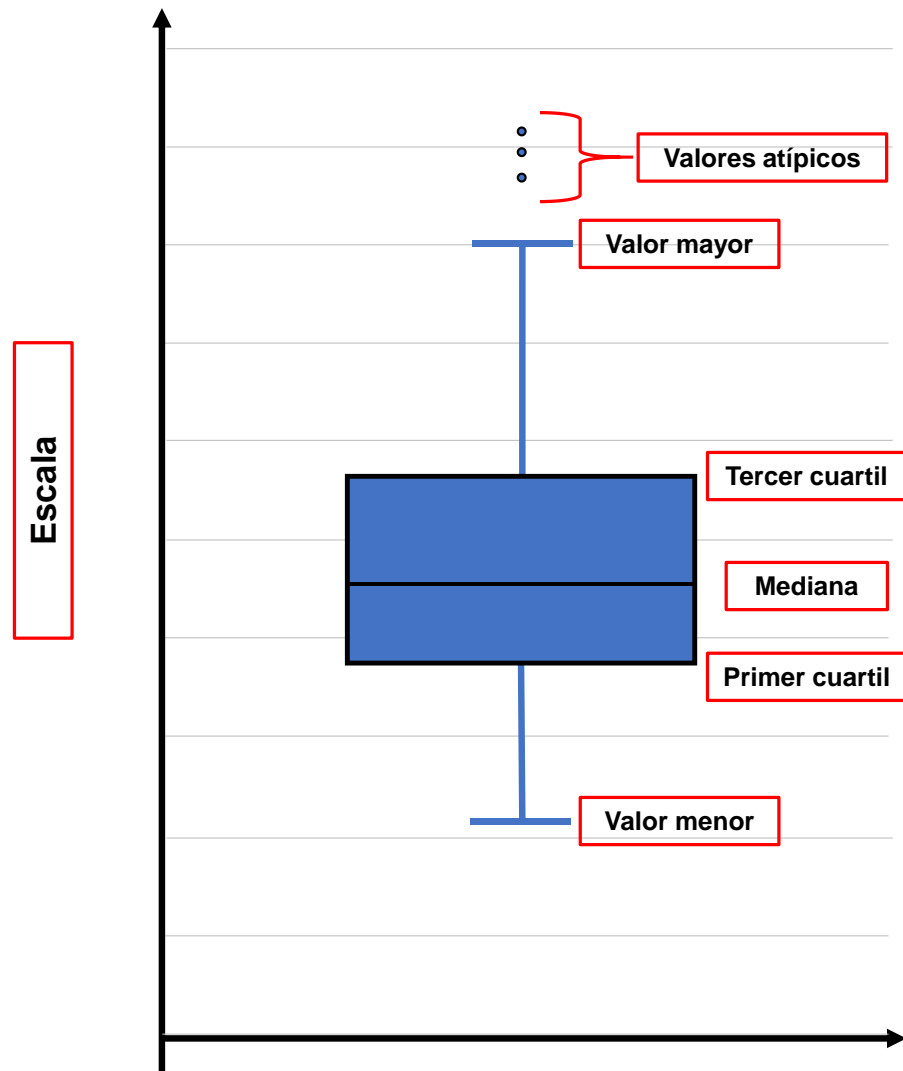


Figura 2.2 Diagrama de Cajas y Bigotes (Boxplot).

2.3 Prueba gamma.

La medida Gamma fue propuesta por Goodman y Kruskal en 1954 (Goodman & Kruskal, 1954) y su valor se define como:

$$\gamma = \frac{P_c - P_d}{P_c + P_d}, \quad \dots\dots\dots (2.2)$$

que es la diferencia entre las probabilidades de concordancia y discordancia para aquellos pares no ligados por alguna variable. Entendiendo que un par es concordante si la observación que tiene una clasificación más alta en la variable x también se clasifica más alto en la variable y ; y discordante si la observación que tiene una clasificación más alta en la variable x se clasifica más bajo en y .

Su rango de variación es de 0 hasta 1. Si las variables son independientes, entonces $P_c = P_d$ lo que implica que $\gamma = 0$, de lo contrario si la relación entre las variables es perfecta y positiva $P_c - P_d = P_c + P_d$ y por lo tanto $\gamma = 1$ y se considera que las dos variables tienen relación una con otra.

2.4 Validación Cruzada (k-fold Cross-validation)

Obtener un buen estimado del error derivado del modelo es de gran importancia, esto puede lograrse al separar los ejemplos de entrada de la base en dos partes: un conjunto de entrenamiento y uno de prueba, este último es usado para evaluar la hipótesis del modelo. Dada la naturaleza estocástica al construir estos conjuntos de datos aunada a la de los datos mismos, no necesariamente se obtendrán resultados idénticos y, en consecuencia, el error obtenido de cada modelo varía (Anthony & Holden, 1998).

La validación cruzada consiste en tomar los datos originales y crear a partir de estos k subconjuntos, los cuales sirven con un doble propósito, ser conjunto de entrenamiento y de prueba. Por lo tanto, este método se repite k veces, de tal forma que cada vez uno de los subconjuntos es utilizado como conjunto de prueba y los otros $k - 1$ subconjuntos son agrupados para conformar el conjunto de entrenamiento. El error estimado en cada caso es promediado sobre todos los k casos para así obtener la efectividad total del modelo. Como es evidente, cada dato del conjunto total sirve como validación y a su vez se encuentra $k - 1$ veces en el conjunto de entrenamiento, lo que reduce significativamente el posible sesgo encontrado cuando se utiliza un alto porcentaje de la

información disponible para entrenar los modelos, de igual forma, reduce la varianza ya que la mayor parte de la información es usada para el proceso de validación (Russell & Norvig, 2010). Como una regla general, k toma valores de 5 o 10, al costo de tener tiempos de computo 5 o 10 veces mayores.

2.5 Árboles de decisión.

Los árboles de clasificación y regresión (CART: Classification and Regression Tree) son una técnica de aprendizaje de árboles de decisión no paramétrica que produce árboles de regresión en caso de que la variable sea numérica; y de clasificación en caso de que la variable sea categórica. El algoritmo CART (Breiman, 1984) tiene como resultado general un árbol donde las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación, también conocida como partición, formando así grupos homogéneos mutuamente excluyentes respecto a la variable a discriminar. Los árboles se construyen mediante un algoritmo de segmentación recursiva deteniéndose una vez se alcanza un criterio de parada. El método utiliza datos históricos para construir el árbol y una vez terminado puede ser usado para clasificar nuevos datos.

Si Y es una variable respuesta y las p variables predictoras son x_1, x_2, \dots, x_p , donde las x se consideran fijas y Y es una variable aleatoria, el problema estadístico consiste en establecer una relación entre Y y las x de tal forma que sea posible predecir Y basado en los valores de x . Matemáticamente se requiere estudiar la probabilidad condicional de la variable aleatoria. $P [Y = y | x_1, x_2, \dots, x_p]$ o una función de su probabilidad tal como la esperanza condicional.

2.5.1 Elementos de un Árbol

El árbol de la **Figura 2.3**, se compone de un nodo inicial llamado nodo raíz formado a partir del atributo que crea el subconjunto más puro, éste se divide a su vez en dos grupos o nodos de decisión, para posteriormente aplicar a estos por separado el procedimiento de partición. Las divisiones se seleccionan de tal forma que la pureza de los nodos de decisión sea mayor al nodo raíz. El objetivo es particionar la respuesta en grupos homogéneos y a la vez mantener el árbol lo suficientemente pequeño.

El proceso de segmentación recursiva continúa hasta que el árbol sea saturado en el sentido de que los sujetos en los nodos descendientes no se pueden partir en una división adicional ya que no cumplen que la pureza de estos nodos sea menor que aquella del nodo de donde provienen. A estos nodos que no pueden seguir siendo divididos se les conoce como nodos terminales.

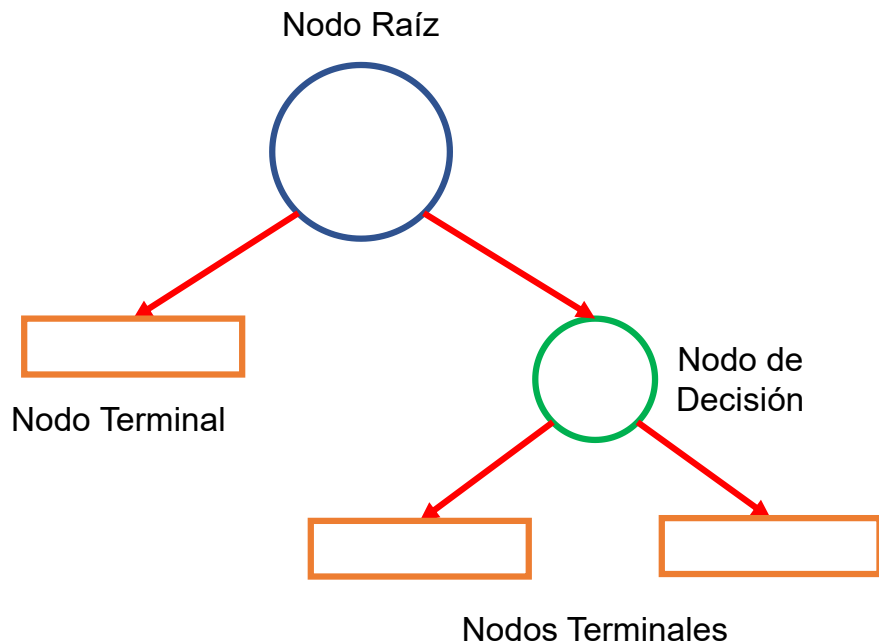


Figura 2.3. Árbol de Decisión.

La metodología para construir y analizar los árboles de regresión y clasificación generalmente consiste en tres pasos (Timofeev, 2004).

1. Construcción del árbol.
2. Poda del árbol.
3. Selección del árbol óptimo mediante un procedimiento de validación cruzada.

2.5.2 Función de Impureza del Nodo

Esta función es una medida que permite determinar la calidad de un nodo y se expresa con $i(t)$. Las medidas de impureza que permiten analizar distintos tipos de respuesta en los árboles de clasificación (Breiman et al, 1984) son:

- Índice de entropía definido como:

$$i(t) = \sum_j p(j|t) \ln p(j|t). \quad \dots\dots\dots (2.3)$$

Donde el objetivo es encontrar la partición que maximice $\Delta i(t)$ de la ecuación mostrada a continuación:

$$\Delta i(t) = - \sum_{j=1}^k p(j|t) \ln p(j|t). \quad \dots\dots\dots (2.4)$$

Índice de Gini como:

$$i(t) = \sum_{i \neq j} p(j|t)p(i|t). \quad \dots\dots\dots (2.5)$$

Para encontrar la partición que maximice $\Delta i(t)$ en:

$$\Delta i(t) = - \sum_{j=1}^k [p_j(t)]^2. \quad \dots\dots\dots (2.6)$$

Respecto a los árboles de regresión, los algoritmos más comunes se encuentran en función del cálculo de:

- Desviación estándar.
- Varianza, y
- Suma de cuadrados.

2.5.3 Poda del árbol

El árbol obtenido la mayor parte del tiempo está sobre ajustado, por lo que deben cortarse sucesivamente nodos terminales hasta encontrar el tamaño más adecuado. Para resolver este problema, una alternativa, es buscar una serie de árboles anidados de tamaños

decrecientes (De'ath & Fabricius, 2000), cada uno de los cuales es el mejor de todos los árboles de su tamaño. Estos árboles pequeños son comparados para determinar el óptimo. Esta comparación está basada en una función de costo-complejidad, $R_\alpha(T)$.

Para cada árbol T , esta función (Deconick & et al, 2006) se define como:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|, \quad \dots\dots\dots (2.7)$$

donde $R(T)$ es el promedio de la suma de cuadrados entre dos nodos, puede ser la tasa de mala clasificación total o la suma de cuadrados residuales total, dependiendo del tipo de árbol, $|\tilde{T}|$ es la complejidad del árbol, definida como el número total de nodos del subárbol y finalmente, α es el parámetro de complejidad.

El parámetro α es un número real mayor o igual a cero, cuando $\alpha = 0$ se tiene el árbol más grande y a medida que este número incrementa, se reduce el tamaño del árbol.

La función $R_\alpha(T)$ siempre será minimizada por el árbol más grande, por tanto, se necesitan mejores estimaciones del error.

De la secuencia de árboles anidados es necesario seleccionar aquél que sea óptimo, y para lograrlo no es efectivo utilizar comparación o penalización de la complejidad (De'ath & Fabricius, 2000), por lo tanto, se requiere estimar con precisión el error de predicción y en general esta estimación se hace utilizando un procedimiento de validación cruzada.

El objetivo es encontrar la proporción óptima entre la tasa de mala clasificación y la complejidad del árbol, siendo la tasa de mala clasificación el cociente entre las observaciones mal clasificadas y el número total de observaciones. El procedimiento de validación cruzada puede implementarse de dos formas, la primera cuando se cuenta con datos suficientes y la segunda cuando no:

- Si se cuenta con datos suficientes, la muestra es partida; sacando la mitad o menos de los datos, posteriormente la secuencia de árboles es construida utilizando los datos que permanecen, para después predecir para cada árbol la respuesta de los datos que se obtuvieron al iniciar el proceso y calcular el error de

las predicciones, para finalmente seleccionar el árbol con el menor error de predicción.

- Si no se cuenta con suficientes datos se aplica la herramienta de validación cruzada con k particiones (k -fold cross-validation).

2.5.4 Regresión logística o Discriminante Logístico

Cuando se desea clasificar un sujeto dentro de uno o más grupos previamente determinados a partir de un conjunto de características observadas del sujeto, es acertado pensar en la utilización de una medida probabilística.

La regresión logística estima la probabilidad de un suceso en función de un conjunto de variables explicativas y en la construcción del modelo no hay ningún supuesto en cuanto a la distribución de probabilidad de las variables por lo que puede incluirse cualquier tipo de variable.

Este modelo puede considerarse como una fórmula para calcular la probabilidad de pertenencia a uno de los grupos, de manera que estima la probabilidad de que una observación pertenezca a alguno. El modelo de regresión logística se formula matemáticamente relacionando la probabilidad de ocurrencia de algún evento E , condicionado a un vector x , de variables explicativas, a través de la forma funcional logística (Press & Wilson, 1978). Así se llega a la siguiente formulación.

$$p(x) = P(E|x) = \frac{1}{1 + e^{-\alpha - \beta^T x}}, \quad \dots \dots \dots (2.8)$$

donde $(\alpha \text{ y } \beta)$ son parámetros desconocidos que se estiman de los datos.

Este modelo puede usarse para clasificar un objeto en una de las poblaciones, siendo E el evento que el objeto pertenezca a la primera población, y x denote un vector de atributos del objeto que será clasificado.

Una medida útil para verificar la calidad de en las clasificaciones obtenidas lo puede ser la tasa de desaciertos, que es la proporción de observaciones mal clasificadas. El modelo

de regresión logística tiene como ventaja que es claro y pueden usarse todos los tipos de variables.

2.5.5 Regresión Logística Multinomial

Esta técnica consiste en la estimación de la probabilidad de que una observación x pertenezca a uno de los grupos, dados valores de las p variables que la conforman.

El modelo compara $G - 1$ categorías contra una categoría de referencia. Dadas n observaciones (y_i, x_i) donde x_i es un vector con p variables y y_i es una variable aleatoria independiente multinomial con valores $1, 2, \dots, G$ la cual indica el grupo al que pertenece cada observación, la probabilidad condicional de pertenencia de x_i a cada grupo está dada por:

$$P(y = j|x_i) = \frac{\exp(\alpha_{1j} + \beta'_{1j}x_i)}{1 + \sum_{k=2}^G \exp(\alpha_{1k} + \beta'_{1k}x_i)} \quad \dots \dots \dots (2.9)$$

donde $\alpha_{11} = \beta_{11} = 0$.

Para clasificar la observación p -variada, en un grupo, se calcula la probabilidad de pertenencia a cada uno de los G grupos y se asigna la mayor probabilidad (Hosmer & Lemeshow, 1989).

2.5.6 Algoritmo M5P para Árboles de Regresión

M5P es una reconstrucción del algoritmo M5 (Quinlan, 1992) para la generación de árboles con modelos de regresión, donde a partir de los valores de los atributos es posible calcular el valor estimado de la instancia.

Este algoritmo asigna una función de regresión lineal a los nodos terminales y ajusta un modelo de regresión multivariable a cada subespacio una vez el espacio completo de datos ha sido clasificado o dividido.

El método M5P para árboles de regresión puede manejar tareas con una alta dimensionalidad, además de que logra trabajar con problemas de clases continuas en lugar de discretas. Logra revelar información de gran importancia de los subconjuntos construidos para cada uno de los modelos lineales para así aproximar relaciones no-lineales del conjunto de datos.

Técnicas diseñadas por Breiman (Breiman et al, 1984) para su sistema CART fueron adaptados para hacer frente a atributos enumerados y valores faltantes. Todos los atributos enumerados son convertidos en variables binarias para que así solo pueda haber dos divisiones cada que sea requerido, donde el criterio de separación está basado en el cálculo del error de cada nodo, éste es después analizado mediante la desviación estándar de los valores de clase que arriban a cada nodo.

De las muchas ventajas que árboles de decisión tienen sobre las técnicas tradicionales de datos multivariantes, sobresalen las siguientes cuatro:

- Comparado con otros algoritmos, los árboles de decisión requieren menos esfuerzo en la preparación de datos durante el pre-procesado.
- No se requiere la normalización de los datos para su uso en un algoritmo CART.
- Valores faltantes en los datos no afectan el proceso de construcción del árbol de decisión de forma considerable.
- El modelo de un árbol de decisión es muy intuitivo y su explicación hacia un grupo de trabajo suele ser sencilla.

2.6 Análisis Factorial.

El análisis factorial es una herramienta estadística multivariante de reducción de datos que es útil para encontrar grupos homogéneos de variables, conocidos como factores, hallados en una matriz de datos que está formada por una gran cantidad de variables (Carles, 2008).

Los grupos homogéneos se forman con las variables que se encuentran altamente relacionadas entre sí, procurando inicialmente que dichos grupos sean independientes de otros.

Cuando un fenómeno estudiado está en función de una gran cantidad de propiedades puede existir el interés de averiguar si estas se agrupan de alguna forma característica. Aplicar análisis factorial a la matriz de datos formada por los atributos y las distintas instancias facilita encontrar grupos con un significado en común.

Por lo tanto, esta técnica permite establecer el mínimo número de dimensiones necesarias para poder explicar el fenómeno en cuestión, explicando a su vez, la mayor parte de la información contenida en los datos.

Fundamentalmente lo que se pretende con el análisis de componentes principales, aquí utilizado, es simplificar la información que aporta una matriz de datos con la finalidad de interpretarla más fácilmente.

Ahora bien, la variabilidad de los datos que se busca explicar es conocida como varianza, la que puede ser definida como la medida de dispersión de una variable aleatoria, es decir, que tan lejos se encuentran los datos de su media, y se calcula con la siguiente fórmula (Fisher, 1918).

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \bar{X})^2}{N - 1}, \quad \dots\dots\dots (2.10)$$

donde

$\sigma^2 = \text{Varianza}$.

$x_i = \text{Valor del Dato } i \text{ de la Variable}$.

$\bar{X} = \text{Media o Promedio Aritmético de los datos de la Variable}$.

Un factor puede entenderse como una suma de respuestas a una serie de variables, una combinación lineal de variables (**variable a + variable b + ... variable n**). La suma total de variables es distinta para cada instancia, la varianza de los totales expresa la diversidad que existe entre cada instancia.

Teniendo, por ejemplo, tres factores, esto quiere decir que es posible descomponer la matriz de datos original en tres grupos; cada uno compuesto por todas las variables, pero en cada grupo las variables tienen un peso específico distinto según sea su relación con cada factor.

Los pesos pueden ser grandes o pequeños, positivos o negativos. Generalmente, en cada factor hay variables con pesos grandes y otros próximos a cero; las variables que

más pesen en cada factor, que tengan un número de correlación con el factor mayor a 0.5 y ningún otro factor tenga una correlación mayor con dicha variable, son aquellas que definen al factor. La varianza de todas las nuevas medidas equivale a la varianza máxima explicable de la medida original.

Existen ciertas restricciones y consideraciones a cumplir para poder hacer uso del análisis factorial, estas corresponden tanto a las instancias como a las variables de la matriz de datos.

En cuanto al número de instancias existe cierta discrepancia entre cuál es el número óptimo para hacer uso del análisis factorial, algunas de estas consideraciones se mencionan a continuación:

- El número de instancias debe ser mayor al doble del número de variables.
- El número de instancias debe ser mayor a cuatro o cinco veces el número de variables.
- El número de instancias debe ser mayor a 10 veces el número de variables.

Si bien es cierto que hay que cumplir con alguna de estas condiciones, dependerá del modelador y de la cantidad y calidad de los datos disponibles y el error aceptable en el modelo factorial, ya que, a mayor número de instancias menor será el error típico de los coeficientes de correlación, por lo tanto, es menor la probabilidad de obtener factores basados en la casualidad, donde la finalidad de esta herramienta es obtener factores causales.


Ahora bien, respecto a las variables, estas deben cumplir con tres condiciones.

- Su distribución debe ser normal.
- Debe haber homocedasticidad, es decir, la varianza de los errores no cambia en el tiempo, y
- debe haber relación entre variables, conocida también como multicolinealidad.

Con la finalidad de determinar si se cumple con lo anteriormente descrito existen ciertas pruebas a aplicar a la matriz de datos, la primera se conoce como prueba Kaiser-Meyer-Olkin (KMO) (Kaiser, 1974), la cual tiene como premisa que, si las variables se encuentran en un mismo factor, sus correlaciones parciales, entendidas también como correlaciones entre pares de variables, deberían ser bajas.

Los valores resultantes de esta prueba se encuentran en un rango entre 0 y 1, donde para valores mayores o iguales a 0.6, se entiende que las correlaciones entre todas las variables analizadas son fuertes, mientras que las correlaciones parciales entre pares de variables son pequeñas, por otro lado, si el resultado de la prueba KMO es menor a 0.6, las correlaciones entre variables son débiles y existe una predisposición entre pares de variables a relacionarse entre sí, pero no con otras.

La siguiente prueba por realizar es conocida como Prueba de Esfericidad de Bartlett, esta es aplicada a la matriz de correlaciones entre variables resultante del análisis de la matriz de datos original, y busca comparar si la matriz de correlaciones observada se ajusta o no a una matriz identidad. Para confirmar que la matriz de datos pasó esta segunda prueba, la matriz de correlaciones debe ser lo suficientemente distinta a la matriz identidad, por lo que se busca que el p-valor total obtenido sea menor a 0.05. Lo anteriormente descrito se ejemplifica en la **Figura 2.4**.



$$[A] = \begin{bmatrix} 1 & 0.5 & 0.9 & 0.34 & 0.41 \\ 0.5 & 1 & 0.4 & 0.46 & 0.41 \\ 0.9 & 0.4 & 1 & 0.55 & 0.12 \\ 0.34 & 0.55 & 0.46 & 1 & 0.54 \\ 0.41 & 0.44 & 0.12 & 0.54 & 1 \end{bmatrix} \neq [A] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Matriz de Correlaciones Observada Matriz Identidad

Figura 2.4. Prueba de Esfericidad de Bartlett.

Para esta última prueba es importante considerar que el resultado obtenido es sensible al número de instancias con el que se está trabajando. De cumplirse lo anterior, el análisis

La ecuación matricial del modelo presentado en la ecuación 2.12 se expresa como sigue:

$$\begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_p \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} & \dots & I_{1m} \\ I_{21} & I_{22} & \dots & I_{2m} \\ \dots & \dots & \ddots & \dots \\ I_{p1} & I_{p2} & \dots & I_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \dots \\ F_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_p \end{bmatrix}, \quad \dots \dots \dots \quad (2.12)$$

en forma matricial $\mathbf{X} = \mathbf{LF} + \mathbf{e}$.

Hipótesis sobre los Factores Comunes

- La esperanza de cada uno de los factores comunes es nula $E(\mathbf{f}) = \mathbf{0}$.
- La matriz de covarianzas de los factores comunes es la matriz identidad $E(\mathbf{f}\mathbf{f}') = \mathbf{I}$

En consecuencia, los factores comunes no están correlacionados entre sí, ya que todos los elementos que no se encuentran en la diagonal principal son nulos.

Así pues, los factores comunes son variables tipificadas de media cero y varianza 1, además, no están correlacionados entre sí.

Hipótesis sobre los Factores Únicos

- La esperanza de cada uno de los factores únicos es nula $E(\mathbf{e}) = \mathbf{0}$.
- La matriz de covarianzas de los factores comunes es la matriz diagonal $\mathbf{\Omega}$, $E(\mathbf{e}\mathbf{e}') = \mathbf{\Omega}$

Por tanto, las varianzas de los factores únicos pueden ser distintas, además, los factores únicos están correlacionados entre sí.

- La matriz de covarianzas entre los factores comunes y los factores únicos es la matriz nula $E(\mathbf{f}\mathbf{e}') = \mathbf{0}$.

Para poder hacer inferencias para cada variable que permitan distinguir entre los factores comunes y el factor único es necesario postular que, *los factores comunes están intercorrelacionados con el factor común único*.

Métodos para la Extracción de Factores

A continuación, se explica el método de extracción de factores utilizando la técnica de *componentes principales* (Hotelling, 1933).

Se parte del modelo factorial mostrado en la ecuación 2.11, considerando las variables observables tipificadas (X_1, X_2, \dots, X_p) . La varianza total de las P variables X_j será P .

De este total, la varianza explicada por los factores comunes es la suma de las comunalidades, y la explicada exclusivamente por el factor F_j es:

$$V_j = I_{1j}^2 + I_{2j}^2 + \dots + I_{pj}^2. \quad \dots\dots\dots (2.13)$$

El coeficiente de correlación poblacional ρ_{hj} entre las variables X_H y X_j , en función de los factores comunes que comparten, se encuentra dada por la expresión:

$$\rho_{hj} = I_{h1}I_{j1} + I_{h2}I_{j2} + \dots + I_{hm}I_{jm} = \sum_{k=1}^m I_{hk}I_{jk}. \quad \dots\dots\dots (2.14)$$

Pudiendo estimarse el coeficiente de correlación poblacional ρ_{hj} por el coeficiente correlacional muestral r_{hj} .

El método del factor principal obtiene el primer factor maximizando la varianza explicada por él, que es:

$$V_1 = I_{11}^2 + I_{21}^2 + \dots + I_{p1}^2, \quad \dots\dots\dots (2.15)$$

sujeto a las restricciones: $r_{hj} = \sum_{k=1}^m I_{hk}I_{jk}$.

El problema de optimización con restricciones se resuelve por el método de los multiplicadores de Lagrange, considerando la función siguiente:

$$G_1 = V_1 + \sum_{h,j=1}^P \vartheta_{hj} \left(r_{hj} - \sum_{k=1}^m I_{hk}I_{jk} \right), \quad \dots\dots\dots (2.16)$$

donde

$\vartheta_{hj} = \text{multiplicadores de Lagrange.}$

Derivando la función lagrangiana respecto a las incógnitas I_{hk} , e igualando a cero, se obtiene la expresión fundamental:

$$\frac{\partial G_1}{\partial I_{hk}} = \delta_{1k} I_{h1} - \sum_{j=1}^P \vartheta_{hj} = 0. \quad \delta = \begin{cases} 1 & \text{si } k = 1 \\ 0 & \text{si } k \neq 1 \end{cases} \dots\dots\dots (2.17)$$

Para $k = 1$, en esta expresión fundamental se tiene que $I_{h1} = \sum_{j=1}^P \vartheta_{hj}$.

Por otra parte, si en la expresión fundamental se multiplica a ambos lados por I_{h1} y se suma respecto a h , se tiene que:

$$\delta_{1k} \sum_{h=1}^P I_{h1}^2 - \sum_{j=1}^P \sum_{h=1}^P \vartheta_{hj} I_{h1} I_{jk} = 0, \quad \delta_{1k} = \begin{cases} 1 & \text{si } k = 1 \\ 0 & \text{si } k \neq 1 \end{cases} \dots\dots\dots (2.18)$$

Denominando $\lambda_1 = \sum_{h=1}^P I_{h1}^2$, teniendo en cuenta que $I_{h1} = \sum_{j=1}^P \vartheta_{hj} I_{j1} \rightarrow I_{j1} = \sum_{h=1}^P \vartheta_{hj} I_{h1}$ ($\vartheta_{hj} = \vartheta_{jh}$).

Se puede escribir:

$$\delta_{1k} \sum_{h=1}^P I_{h1}^2 - \sum_{j=1}^P \sum_{h=1}^P \vartheta_{hj} I_{h1} I_{jk} = 0 \rightarrow \delta_{1k} \lambda_1 - \sum_{j=1}^P I_{j1} I_{jk} = 0, \quad \delta_{1k} = \begin{cases} 1 & \text{si } k = 1 \\ 0 & \text{si } k \neq 0 \end{cases} \dots\dots\dots (2.19)$$

Multiplicando la expresión $\delta_{1k} \lambda_1 - \sum_{j=1}^P I_{j1} I_{jk}$ por I_{hk} y sumando en k se obtiene:

$$I_{h1} \lambda_1 - \sum_{j=1}^P I_{j1} \left(\sum_{k=1}^P I_{hk} I_{jk} \right) = 0 \rightarrow \sum_{j=1}^P I_{j1} r_{hj} - i_{h1} \lambda_1 = 0 \quad h = 1, \dots, P. \quad \dots (2.20)$$

De la expresión $\sum_{j=1}^P I_{j1} r_{hj} - i_{h1} \lambda_1 = 0$, se obtiene:

$$\begin{cases} (h_1^2 - \lambda_1)I_{11} + r_{12}I_{21} + \dots + r_{1p}I_{p1} = 0 \\ r_{21}I_{11} + (h_2^2 - \lambda_1)I_{21} + \dots + r_{2p}I_{p1} = 0, \dots\dots\dots (2.21) \\ \dots\dots\dots \\ r_{p1}I_{11} + r_{p2}I_{21} + \dots + (h_n^2 - \lambda_1)\alpha_{n1} = 0 \end{cases}$$

λ_1 es el mayor valor propio de la matriz de correlaciones LL' y $(I_{11}, I_{21}, \dots, I_{p1})'$ es su vector propio asociado, de módulo λ_1 . En consecuencia, se tiene que:

$$I_{i1} = \alpha_{i1}\sqrt{\lambda_1} \quad i = 1,2, \dots, P, \dots\dots\dots (2.22)$$

Siendo $(\alpha_{11}, \alpha_{21}, \dots, \alpha_{p1})$ un vector propio de módulo unidad.

Una vez obtenidos los pesos (cargas factoriales) del primer factor, que es el que más contribuye a la varianza de las variables, se elimina su influencia considerando el nuevo modelo factorial como:

$$\begin{cases} X'_1 = I_{11}F_1 + I_{12}F_2 + \dots + I_{1m}F_m + e_1 \\ X'_2 = I_{21}F_1 + I_{22}F_2 + \dots + I_{2m}F_m + e_2, \dots\dots\dots (2.23) \\ \dots\dots\dots \\ X'_p = I_{p1}F_1 + I_{p2}F_2 + \dots + I_{pm}F_m + e_p \end{cases}$$

Se obtiene el segundo factor maximizando la varianza explicada por éste en el segundo modelo, que es:

$$V_2 = I_{12}^2 + I_{22}^2 + \dots + I_{p2}^2, \dots\dots\dots (2.24)$$

sujeto a las restricciones: $r_{hj} = \sum_{k=1}^P I_{hk}I_{jk}$.

Como se realizó anteriormente, se demuestra que:

$$I_{i2} = \alpha_{i2}\sqrt{\lambda_2} \quad i = 1,2, \dots, P. \dots\dots\dots (2.25)$$

Siendo $(\alpha_{11}, \alpha_{21}, \dots, \alpha_{p1})$ un vector propio de módulo unidad y λ_2 el segundo mayor valor propio de la matriz de correlaciones LL' .

Este proceso se sigue repitiendo hasta obtener los pesos o cargas factoriales de todos los factores, esto es, la matriz factorial, o al menos hasta que la varianza total explicada por los factores comunes sea igual o próxima a la suma de las comunalidades.

El número de factores obtenidos coincide con el de los valores propios no nulos de LL' , que todos son positivos ya que LL' es simétrica semidefinida positiva.

El método del factor principal puede explicarse por la diagonalización de la matriz LL' , que tomará la forma siguiente:

$$LL' = TD_{\lambda}T', \quad \dots\dots\dots (2.26)$$

donde:

$T \equiv$ matriz cuyas k columnas son los vectores propios de módulo de unidad LL' .

$D_{\lambda} \equiv$ diagonal $(\lambda_1 \dots \lambda_k)$.

La matriz factorial será entonces $\rightarrow L = TD_{\lambda}^{1/2}$.

Contrastes en el Modelo Factorial

En el modelo factorial pueden realizarse varios tipos de contrastes, que suelen agruparse en dos tipos de bloques, según se apliquen previamente a la extracción de los factores o después a esta.

Con los contrastes aplicados previamente a la extracción de factores, como es el caso de este trabajo, se trata de analizar la pertenencia de la aplicación del análisis factorial a un conjunto de variables observables.

Entre los contrastes que aplican se encuentra: Esfericidad de Bartlett y la medida de adecuación muestral de Kaiser-Meyer-Olkin.

- **Contraste de Esfericidad de Bartlett:** Antes de realizar un análisis factorial se plantea si están correlacionadas entre sí las variables o atributos originales. Si no

lo estuvieran no existirían factores comunes, y, por lo tanto, no tendría sentido aplicar el análisis factorial. Esta cuestión suele probarse utilizando el contraste de la esfericidad de Bartlett.

La matriz de correlación poblacional R_P recoge la relación entre cada par de variables mediante sus elementos ρ_{ij} situados fuera de la diagonal principal. Los elementos de la diagonal principal son unos, dado que toda variable está totalmente relacionada consigo misma.

En el supuesto de que no existiese ninguna relación entre las P variables en estudio, la matriz R_P sería la identidad, cuyo determinante es la unidad.

En consecuencia, para decidir la ausencia o no de la relación entre las P variables puede plantearse el siguiente contraste:

$$H_0: |R_P| = 1 \quad H_1: |R_P| \neq 1.$$

La hipótesis nula H_0 a contrastar es que todos los coeficientes de correlación teóricos entre cada par de variable son nulos.

Bartlett introdujo un estadístico para este contraste, basado en la matriz de correlación muestral R , que bajo la hipótesis nula H_0 , tiene una distribución chi-cuadrado con $\frac{p(p-1)}{2}$ grados de libertad.

$$\chi^2_{0.5(k^2-k)} = - \left[n - 1 - \frac{1}{6}(2k + 5) \right] \ln |R|. \quad \dots\dots\dots (2.27)$$

- **Medida de Adecuación Muestral Global (KMO):** Los estadísticos Kaiser, Meyer y Olkin propusieron una medida de adecuación muestral al análisis factorial, que es conocida por las iniciales KMO.

En un modelo con distintas variables el coeficiente de correlación parcial entre dos variables mide la correlación existente entre ellas, una vez que se han descontado los efectos lineales del resto de las variables del modelo. En este modelo factorial se pueden considerar esos efectos de otras variables como los correspondientes a los factores comunes.

En consecuencia, el coeficiente de correlación parcial entre dos variables sería equivalente al coeficiente de correlación entre los factores únicos de esas dos variables.

De acuerdo con el modelo de análisis factorial, los coeficientes de correlación teóricos calculados entre cada par de factores únicos son nulos por hipótesis. Si los coeficientes de correlación parcial constituyen una aproximación a dichos coeficientes teóricos, deben estar próximos a cero. La medida KMO se define mediante la expresión:

$$KMO = \frac{\sum_j \sum_{h \neq j} r_{jh}^2}{\sum_j \sum_{h \neq j} r_{jh}^2 + \sum_j \sum_{h \neq j} a_{jh}^2}, \quad \dots \quad (2.28)$$

donde

$r_{jh} \equiv$ coeficientes de correlación observados entre las variables X_j y X_h .

$a_{jh} \equiv$ coeficiente de correlación parcial entre las variables X_j y X_h .

En caso de que exista adecuación de los datos a un modelo de análisis factorial, el término del denominador que recoge los coeficientes a_{jh} será pequeño y, en consecuencia, la medida KMO será próxima a la unidad.

Valores de KMO por debajo de 0.6 no serán aceptables, considerándose inadecuados los datos a un modelo de análisis factorial.

Para valores de KMO mayores a 0.6 se considera aceptable la adecuación de los datos a un modelo de análisis factorial. Mientras más cerca están de uno los valores de KMO, mejor es la adecuación de los datos a un modelo factorial, considerándose ya excelente para los valores próximos a 0.9.

Rotación de los Factores

El trabajo en el análisis factorial persigue que los factores comunes tengan una interpretación clara, porque de esa forma se analizan mejor las interrelaciones existentes entre las variables originales. Sin embargo, en muy pocas ocasiones resulta fácil

encontrar una interpretación adecuada de los factores iniciales, con independencia del método que se hubiera utilizado para su extracción.

Los procedimientos de rotación de factores se han ideado para obtener, a partir de la solución inicial, unos factores que sean fácilmente interpretables.

En la solución inicial cada uno de los factores comunes está correlacionado en mayor o menor medida con cada una de las variables originales.

Con los factores rotados se trata de que cada una de las variables originales tenga una correlación lo más próxima a uno que sea posible, con uno de los factores y correlaciones próximas a cero con el resto de los factores.

Dado que hay más variables que factores comunes, cada factor tendrá una correlación alta con un grupo de variables y baja con el resto.

Examinando las características de las variables de un grupo asociado a un determinado factor se pueden encontrar rasgos comunes que permiten identificar el factor y darle una denominación que responda a esos rasgos comunes.

Si se consigue identificar claramente estos rasgos, se habrá dado un paso importante, ya que con los factores comunes no sólo se reducirá la dimensión del problema, sino que también se conseguirá desvelar la naturaleza entre las interrelaciones existentes entre las variables originales.

Existen dos formas básicas de realizar la rotación de factores, la rotación ortogonal y la rotación oblicua. Dados los alcances de este trabajo, la explicación se centrará en la rotación ortogonal, más particularmente la rotación varimax.

- **Método de Rotación Ortogonal: Rotación Varimax.**

Este método obtiene los ejes de los factores maximizando la suma de varianzas de las cargas factoriales al cuadrado dentro de cada factor.

La varianza de las cargas factoriales al cuadrado del factor i -ésimo se puede calcular a partir de los momentos respecto al origen como sigue:

$$S_i^2 = \frac{\sum_{j=1}^P (I_{ji}^2)^2}{P} - \left(\frac{\sum_{j=1}^P I_{ji}^2}{P} \right)^2. \quad \dots\dots\dots (2.29)$$

La suma de varianzas de las cargas factoriales al cuadrado dentro de cada factor, con m factores seleccionados será:

$$S^2 = \sum_{i=1}^m S_i^2 = \sum_{i=1}^m \left[\frac{\sum_{j=1}^P (I_{ji}^2)^2}{P} - \left(\frac{\sum_{j=1}^P I_{ji}^2}{P} \right)^2 \right]. \quad \dots\dots\dots (2.30)$$

La expresión 2.30 plantea el problema que las variables con mayores comunalidades tienen una mayor influencia en la solución final. Para evitarlo se efectúa la normalización de Kaiser, en la que cada carga factorial al cuadrado se divide por la comunalidad de la variable correspondiente.

Cuando se aplica la normalización de Kaiser, el método recibe el nombre de Varimax normalizado. En esta línea, la expresión que se maximiza es:

$$SN^2 = \sum_{i=1}^m \left[\frac{\sum_{j=1}^P \left(\frac{I_{ji}^2}{h_j^2} \right)^2}{P} - \left(\frac{\sum_{j=1}^P \frac{I_{ji}^2}{h_j^2}}{P} \right)^2 \right] = \sum_{i=1}^m \left[\frac{1}{P} \sum_{j=1}^P \left(\frac{I_{ji}^2}{h_j^2} \right)^2 - \left(\frac{1}{P} \sum_{j=1}^P \frac{I_{ji}^2}{h_j^2} \right)^2 \right]. \quad (2.31)$$

En su forma definitiva, el método varimax halla la matriz B maximizando:

$$W = P^2 SN^2 = \left[P \sum_{j=1}^P \left(\frac{I_{ji}^2}{h_j^2} \right)^2 - \left(\sum_{j=1}^P \frac{I_{ji}^2}{h_j^2} \right)^2 \right] = P \sum_{i=1}^m \sum_{j=1}^P \left(\frac{I_{ji}^2}{h_j^2} \right)^2 - \sum_{i=1}^m \sum_{j=1}^P \left(\sum_{j=1}^P \frac{I_{ji}^2}{h_j^2} \right)^2. \quad (2.32)$$

Para realizar la maximización se halla la matriz $T = \begin{bmatrix} \cos \varphi & \text{sen} \varphi \\ -\text{sen} \varphi & \cos \varphi \end{bmatrix}$, que efectúa la rotación de dos factores de forma que su suma de simplicidades sea máxima.

Repitiendo el proceso para los $\frac{p(p-1)}{2}$ pares de posibles factores se tiene:

$$B = L T_{11} T_{12} T_{13} \dots T_{m-1,m}. \quad \dots \dots \dots (2.33)$$

La sucesión de rotaciones se denomina ciclo. Se repiten los ciclos hasta completar uno que en que todos los ángulos de giro sean menores que un cierto valor prefijado.

Una propiedad importante del método varimax es que, después de aplicarlo, queda inalterada tanto la varianza total explicada por los factores, como la comunalidad de cada una de las variables.

La nueva matriz corresponde también a los factores ortogonales y tiende a simplificar la matriz factorial por columnas, siendo muy adecuada cuando el número de factores es pequeño.

Capítulo 3: Inteligencia Artificial (IA)

La inteligencia artificial es una de las ramas de las ciencias de la computación que más interés ha despertado en la actualidad, debido a su amplio campo de aplicación. Los trabajos en esta área comenzaron formalmente poco después de la Segunda Guerra Mundial, y el nombre como tal fue acuñado en 1956.

En los inicios de la IA existía un conflicto referido a la falta de una definición clara y única, así que no es de sorprender que aún en la actualidad, no exista *una* definición de IA. Las distintas definiciones de la Inteligencia artificial hacen énfasis en diferentes aspectos; aunque existen similitudes entre ellas, algunas se enfocan en el *proceso del pensamiento* mientras que otras lidian con el *comportamiento* (Russell & Norvig, 2010).

A continuación, se presentan algunas de las definiciones más simples:

- Estudios, en computación, que tratan sobre máquinas capaces de percibir, razonar y actuar (Winston, 1992).
- El arte de crear máquinas que realicen funciones que requieren inteligencia cuando las mismas funciones pudieran ser realizadas por una persona (Kurzweil, 1990).
- Nuevo esfuerzo para lograr que la computadora piense... máquinas con mentes, en el sentido completo y literal (Haugeland, 1985).
- Rama de la ciencia computacional preocupada por la automatización de la conducta inteligente (Luger, 1995).
- El estudio de las facultades mentales a través del uso de modelos computacionales (Charniak, Riesbeck, & et al, 1987).
- La IA se ocupa del comportamiento inteligente de los artefactos (Nilsson, 1998).

La inteligencia artificial fue originalmente construida con base en teorías existentes de otras áreas del conocimiento. Algunas de las principales fuentes que nutrieron a esta área son las ciencias de la computación, la lingüística, las matemáticas y la psicología (Ponce Gallegos et al,2014).

Filósofos como Sócrates, Platón, Aristóteles, Leibniz desde el año 400 aC, sentaron las bases para la inteligencia artificial al concebir a la mente como una máquina que funciona a partir del conocimiento codificado en un lenguaje interno y al considerar que el pensamiento servía para determinar cuál era la acción correcta que se debía emprender.

Las Matemáticas proveyeron las herramientas para manipular las aseveraciones de certeza lógica, así como aquellas en las que existe incertidumbre de tipo probabilista; el cálculo por su lado brindó los instrumentos que permiten la modelación de diferentes tipos de fenómenos y fueron también las matemáticas, las que prepararon el terreno para el manejo del razonamiento con algoritmos (Russell & Norvig, 2010).

Las Ciencias de la Computación comenzaron poco antes que la IA misma. Las teorías de la inteligencia artificial encuentran un medio para su implementación de artefactos y modelado cognitivo a través de las computadoras. Los programas de inteligencia artificial por lo general son extensos y no funcionarían sin los avances de velocidad de procesamiento y memoria aportados por la industria del cómputo.

La Economía brindó a la IA una serie de teorías (Teoría de la decisión (North, 1968) – la cual combina la Teoría de la Probabilidad y la Teoría de la Utilidad; Teoría de juegos (Bierman & Fernández, 1998) – para pequeñas economías; Procesos de decisión de Markov – para procesos secuenciales, entre otras) que la posibilitaron para la toma de decisiones.

Finalmente, la Neurociencia ha contribuido a la IA con los conocimientos recabados hasta la fecha sobre la forma en la que el cerebro procesa muchos tipos de información.

Clasificación de la Inteligencia Artificial

La escuela clásica dentro de la IA utiliza representaciones simbólicas basadas en un número finito de primitivas y de reglas para la manipulación de símbolos, por ejemplo, redes semánticas, lógica de predicados, etcétera, los cuales fueron y siguen siendo parte central de dichos sistemas. Otro tipo de representación es el llamado sub-simbólico, el cual utiliza representaciones numéricas o sub-simbólicas del conocimiento. Este enfoque se caracteriza por crear sistemas con capacidad de aprendizaje, el cual puede obtenerse a nivel de individuo imitando el cerebro (Redes Neuronales), o a nivel de especie, imitando la evolución (Algoritmos Genéticos) (Ponce G, et al. 2014).

En la actualidad, la IA empieza a extender sus áreas de investigación en diversas direcciones y trata de integrar diferentes métodos en sistemas de gran escala, en su afán por explotar al máximo las ventajas de cada una de estas en una enorme cantidad de áreas del conocimiento ya que se realizan aplicaciones para medicina, biología, ingeniería, educación, etc.

Existen actualmente nuevas técnicas que utilizan el enfoque sub-simbólico como son los algoritmos de optimización de colonias de hormigas, sistema inmune, cúmulo de partículas, entre otros, los cuales están inspirados en comportamientos emergentes de la naturaleza.

Historia de la Inteligencia Artificial

Fue en los años 50 cuando se logró por primera vez realizar con cierto éxito un sistema de este tipo, se le llamó Perceptrón de Rosenblatt. Este era un sistema visual de reconocimiento de patrones en el cual se aunaron esfuerzos para que se pudieran resolver una gama amplia de problemas.

Aproximadamente en ese tiempo, el matemático inglés Alan Turing (1912-1954) propuso una prueba con la finalidad de demostrar la existencia de la *inteligencia* en un dispositivo no biológico. Esta prueba conocida como “Test de Turing” se fundamenta en la hipótesis de que “si una máquina se comporta en todos aspectos como inteligente, entonces debe serlo” (Alan Turing, 1950).

Posteriormente en 1957, Alan Newell y Herbert Simon, que trabajaban en la demostración de teoremas y el ajedrez por ordenador logran crear un programa llamado GPS (General Problem Solver). Este era un sistema donde el usuario definía un entorno en función de una serie de objetos y los operadores que se podían aplicar sobre ellos. Este programa fue redactado mediante el uso de IPL (Information Processing Lenguaje) y es considerado como el primer programa en el que se separó la información relacionada con el problema de la estrategia empleada para darle solución.

El primer sistema experto fue el denominado Dendral, un intérprete de espectrograma de masa construido en 1967, pero el más influyente resultaría ser el Mycin en 1974, este último era capaz de diagnosticar trastornos en la sangre y recetar la correspondiente medicación.

Por sus implicaciones con áreas como la medicina, psicología, biología, ética y filosofía entre otras, la inteligencia artificial ha tenido que lidiar con fuertes grupos oponentes y críticas desde sus orígenes, sin embargo, siempre existió un grupo de personas interesadas en el área lo que permitió que se consolidara como un área del conocimiento de gran interés para la investigación científica.

Modelos de Inteligencia

Existe una clasificación de modelos de inteligencia artificial que se basa en el objetivo y la forma en que trabaja el sistema, esta clasificación de manera inicial se veía como clases independientes, sin embargo, en la actualidad los sistemas mezclan características de estos modelos (Russell & Norvig, 2010).

- **Sistemas que piensan como humanos:** la aproximación al modelado cognitivo: Si se tiene que decir que un programa piensa como un humano, se debe tener una manera de determinar cómo piensan los humanos. Es necesario adentrarse en el funcionamiento de la mente humana. Existen tres formas de lograrlo, a través de la introspección – tratar de atrapar los pensamientos propios a medida que éstos pasan; ayudándose de experimentos psicológicos – al observar a una persona en acción; y a través de imágenes del cerebro – observando éste en acción. Una vez se tiene una teoría suficientemente precisa de la mente, es posible expresar la teoría como un programa computacional. Si el comportamiento de las entradas y salidas del programa iguala al correspondiente comportamiento humano, es evidencia de que algunos mecanismos del programa podrían también operar en humanos. Concretamente, el modelo es el funcionamiento de la mente humana, pues se intenta establecer una teoría sobre su funcionamiento a partir de la cual se puedan establecer modelos computacionales.

El campo interdisciplinario conocido como *ciencia cognitiva* conjunta a los modelos computacionales de la IA con las técnicas experimentales de la psicología para construir teorías precisas y comprobables sobre la mente humana.

- **Sistemas que piensan racionalmente:** La aproximación de las “Leyes del pensamiento”: El filósofo griego Aristóteles fue uno de los primeros en intentar codificar el “pensamiento correcto”, que es irrefutablemente un proceso de razonamiento. Sus silogismos proveyeron patrones para estructuras argumentales

que siempre llevaban a conclusiones correctas cuando se daban premisas correctas, por ejemplo, “Sócrates es un hombre: todos los hombres son mortales; por lo tanto, Sócrates es mortal”. Estas leyes del pensamiento gobernarían la operación de la mente; sus estudios iniciaron con el área llamada *lógica*. En el siglo XIX los logicistas desarrollaron una notación precisa para declaraciones sobre toda clase de objetos en el mundo y la relación entre ellos. La llamada tradición *logicista* dentro de la inteligencia artificial tiene la esperanza de construir en este tipo de líneas del pensamiento para crear sistemas inteligentes. Sin embargo, existen dos grandes obstáculos en este enfoque. El primero, es referido a que no es fácil tomar conocimiento informal y declararlo en los términos formales requeridos por la notación lógica, particularmente cuando el conocimiento tiene una certeza menor al cien por ciento. El segundo problema es que existe una enorme diferencia entre resolver un problema “en principio” y resolverlo de forma práctica.

- **Sistemas que actúan racionalmente, el enfoque del agente racional:** Un *agente* es conocido como cualquier cosa que puede percibir su ambiente a través de *sensores* y actuar sobre ese ambiente a través de *actuadores*. Por supuesto que todos los programas de computadora hacen algo, pero para los agentes computacionales es esperado que hagan más: operar autónomamente, percibir el ambiente que los rodea, adaptarse a los cambios, y crear y perseguir metas. Un *agente racional* es aquel que actúa de tal forma que pueda lograr el mejor resultado.

El enfoque del agente racional tiene dos ventajas sobre los demás sistemas; primero, es más general que el enfoque de las “leyes del pensamiento”, ya que una correcta inferencia es uno de los muchos mecanismos para lograr racionalidad. La segunda ventaja es respecto a la facilidad que presenta para el desarrollo científico en contraste de aquellos enfoques que se basan en el comportamiento o pensamiento humano. El estándar de la racionalidad se encuentra matemáticamente bien definido y es completamente general. El comportamiento humano, por otro lado, se encuentra bien adaptado para un

ambiente en específico y está definido por la suma total de todas las cosas que los humanos hacen.

- **Sistemas que actúan como humanos, el enfoque de la prueba de Turing:** La prueba de Turing fue diseñada para proveer una definición operacional satisfactoria de inteligencia. Una computadora pasa la prueba si un humano que actúa como interrogador, después de plantear una serie de preguntas, no puede distinguir si las respuestas escritas fueron dadas por una persona o por una computadora. Por lo tanto, una computadora necesita poseer las siguientes capacidades:
 - Procesamiento de lenguaje natural: para permitir una comunicación satisfactoria en cualquier lenguaje humano.
 - Representación del conocimiento: para almacenar toda la información que sepa o haya obtenido.
 - Razonamiento automatizado: para utilizar la información almacenada con la finalidad de responder a las preguntas y llegar a nuevas conclusiones.
 - Aprendizaje de máquina (machine learning): para adaptarse a nuevas circunstancias y detectar y extrapolar patrones.

La prueba de Turing deliberadamente evita contacto físico directo entre el interrogador y la máquina, ya que la simulación física de una persona es innecesaria para la inteligencia (Russell & Norvig, 2010).

Para servir como resumen, la **Figura 3.1**, muestra de forma condensada la información respecto a los modelos de inteligencia.

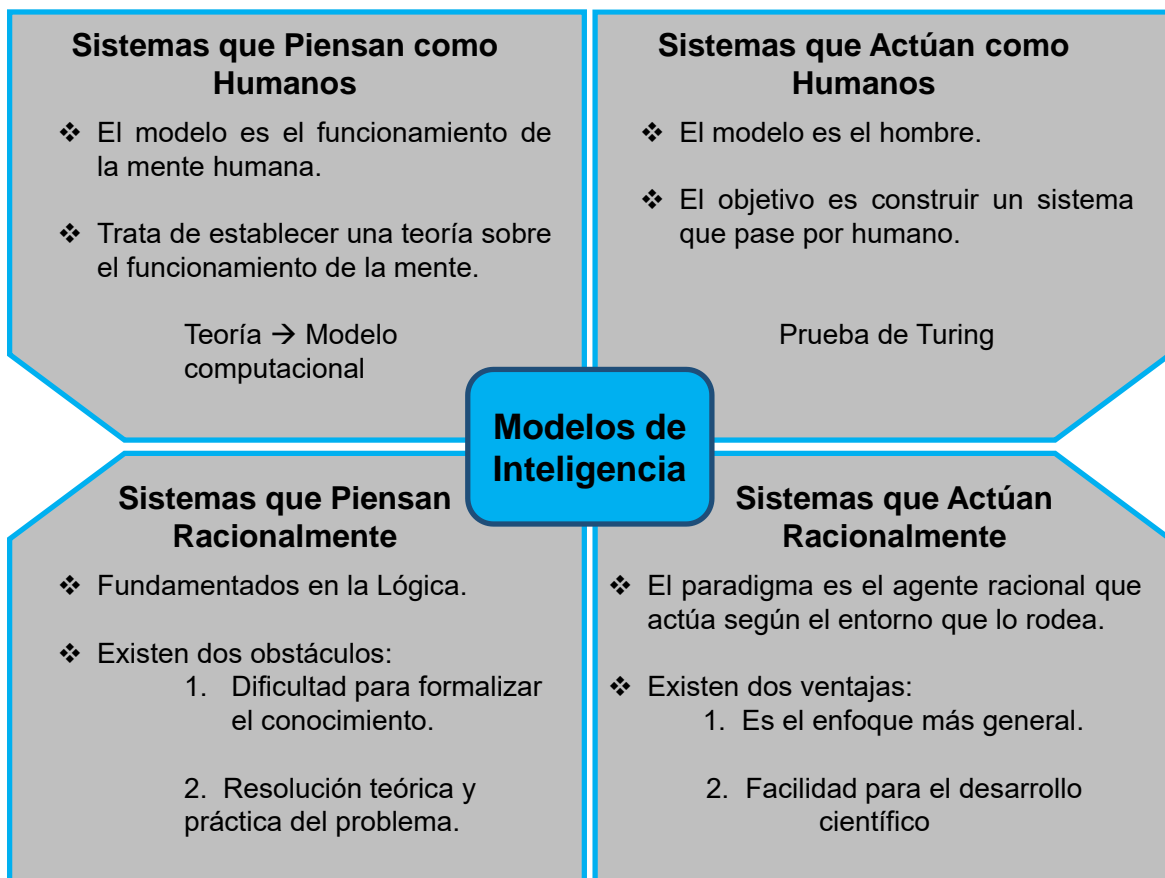


Figura 3.1. Modelos de Inteligencia.

3.1 Redes Neuronales Artificiales (RNA's)

La evolución de las computadoras ha hecho que la tecnología aplicada tenga importantes crecimientos, sobre todo si se tratan problemas con soluciones basados en algoritmos tradicionales, sin embargo, existen aún retos que no pueden ser enfrentados con suficiente asertividad si no se modifican los paradigmas.

Mediante el uso de algoritmos tradicionales (bases de datos, programación orientada a objetos y eventos, etc.) resulta complejo el abordar la solución de problemas donde se requiera descubrir similitudes para clasificación, análisis y reconocimiento de patrones en grandes bases de datos, por ejemplo, y es necesario dar soluciones alternativas. El cambio necesario lo personifica la inteligencia artificial IA, la cual es un intento por descubrir y describir aspectos de la inteligencia humana que pueden simularse mediante el uso de máquinas. Las RNs, como herramienta de la IA, buscan emular el aprendizaje

humano (reconocimiento de patrones, memorización y asociación de hechos entre otros) para ayudar a resolver problemas donde no es posible expresar la solución a través de un algoritmo, y en los que la respuesta al problema requiere de experiencia adquirida.

Una RN consiste en unidades de procesamiento en las que se realiza el intercambio de información. Por su estructuración y sus operaciones, las RN tienen la capacidad de aprender y mejorar su funcionamiento. Las actividades en las que se ha probado su sobresaliente eficiencia van desde análisis de voz, rostro, escritura, hasta extracción de patrones de comportamiento de secuencias multidimensionales dependientes del tiempo (Matich, 2001).

Analogía

Es posible establecer una analogía entre una neurona biológica y una artificial, donde, como muestra la **Figura 3.2**, las conexiones entre los nodos (x_1, x_2, \dots, x_n) representan las dendritas y los axones (entradas), sin embargo, en este caso el axón puede funcionar como una salida (y), si se referencia a una neurona anterior a la que se está analizando. Los pesos (w_1, w_2, \dots, w_n) representan la sinapsis y la aproximación umbral representa la actividad realizada por el núcleo de la célula (soma).

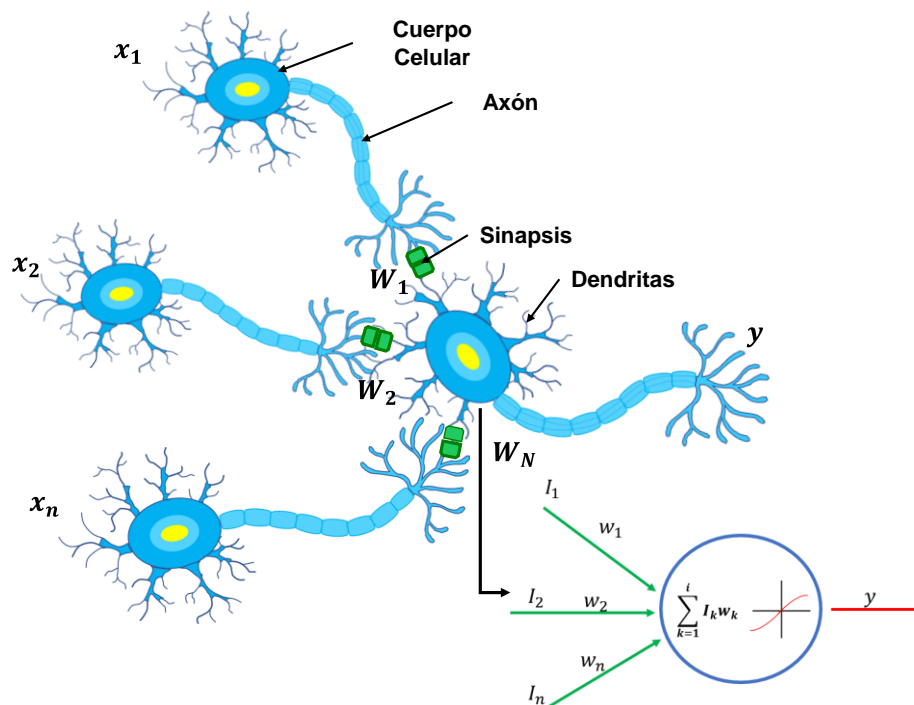


Figura 3.2. Analogía de Neuronas Biológicas con una Neurona Artificial.

3.2 Elementos básicos de una Red Neuronal.

Debido a su naturaleza, existen varias formas de definir una red neuronal, desde las más básicas y genéricas hasta aquellas profundamente detalladas.

De forma general, una red neuronal puede definirse como un modelo computacional que busca simular el proceso de aprendizaje biológico simple (Singh, 2005). Un modelo neuronal consta de una serie de algoritmos numéricos de optimización inspirados en estudios del cerebro y el sistema nervioso (Huang, et al, 1996). Una RN funciona como un sistema de procesamiento (en paralelo, o en semi-paralelismo) de información que es capaz de desarrollar asociaciones y transformaciones entre la información de entrada y la respuesta en la salida (Singh, 2005)

Una red neuronal está caracterizada por dos componentes principalmente, un conjunto de nodos y las conexiones entre éstos. Los nodos pueden verse como unidades computacionales que reciben información externa (inputs) la cual es procesada para obtener una respuesta (output), este proceso descrito puede ser muy simple (tal como sumar las entradas) o bien, muy complejo (donde un nodo podría ser otra red neuronal por si solo) (García S., et al, 2016). Las unidades de procesamiento, por lo tanto, reciben, procesan y transmiten señales tal como las neuronas biológicas.

Ahora bien, las conexiones o pesos (weights) pueden manejar la información de manera unidireccional o bidireccional, y según sea el caso, estos afectan el comportamiento de la red, el cual se entiende como conocimiento emergente. Entre más conexiones se tenga entre los nodos o neuronas, la influencia de estos en el modelo será mayor.

En la **Figura 3.3** se puede observar un esquema de una red neuronal, la cual representa el flujo de la información según el proceso descrito anteriormente, es importante resaltar que en la figura se puede observar una red de tres capas, lo cual puede variar, teniendo una capa de entradas (inputs), una sección llamada capa oculta donde generalmente ocurre el procesamiento de datos, la cual puede estar constituida por más de una capa, y por último una capa de salida (output).

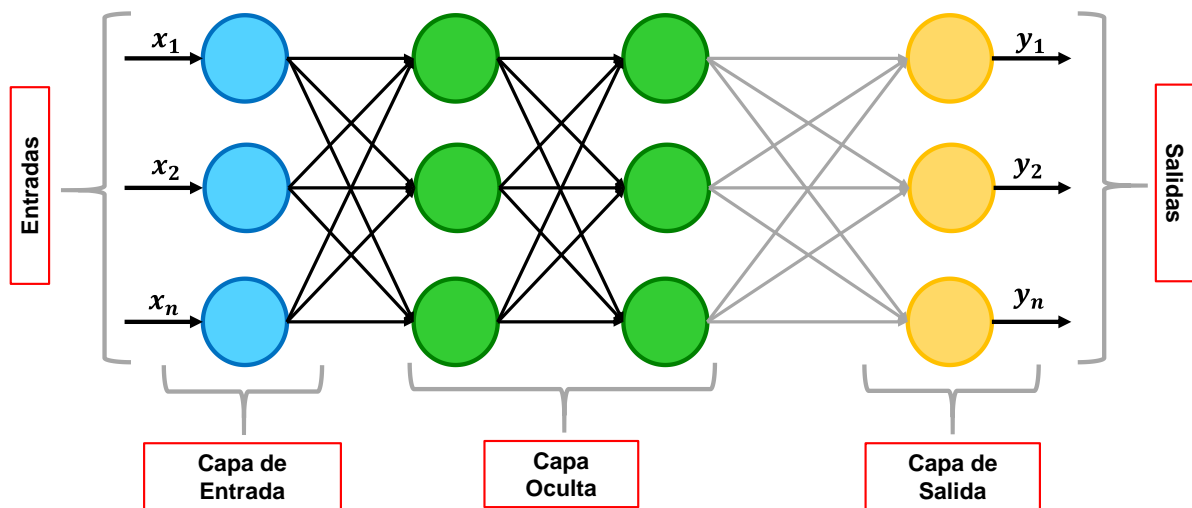


Figura 3.3. Esquematización de una Red Neuronal Artificial multicapa. Modificada de (Matich, 2001).

Las neuronas pueden clasificarse según los valores que puedan tomar, de forma general se distingue entre dos tipos principales:

- Neuronas binarias: estas solamente pueden tomar valores de 0 y 1 o bien de -1 y 1.
- Neuronas reales: estas por otra parte pueden tomar cualquier valor dentro del rango $[0,1]$ o $[-1,1]$.

3.3 Ventajas de las Redes Neuronales

Debido a su constitución y a sus fundamentos. Las redes neuronales artificiales presentan un gran número de características semejantes a las del cerebro. Lo anterior, hace que se ofrezcan numerosas ventajas y que este tipo de tecnología se esté aplicando en múltiples áreas (Matich, 2001). A continuación, se enlistan algunas de estas ventajas importantes a resaltar en trabajo, que consecuentemente serán explicadas en esta sección.

- Aprendizaje Adaptativo.
- Autoorganización.
- Tolerancia a Fallos.
- Operaciones en tiempo Real.

Aprendizaje Adaptativo

Como las redes neuronales pueden aprender a diferenciar patrones mediante ejemplos y entrenamiento, no es necesario especificar funciones de distribución de probabilidad ni modelos que describan algún tipo de fenómeno, comportamiento o relación entre las distintas partes estudiadas.

De lo anterior, se puede entender que las redes neuronales son sistemas dinámicos auto-adaptativos, es decir, tienen una capacidad de ajuste de los elementos procesales (neuronas) que componen el sistema, además, son dinámicos pues son capaces de estar en constante cambio para adaptarse a cualquier nueva condición que se les sea impuesta.

En el proceso de aprendizaje, los enlaces ponderados de las neuronas se ajustan de manera que se obtengan ciertos resultados específicos. Una red neuronal no necesita un algoritmo para resolver un problema, ya que esta puede generar su propia distribución de pesos en los enlaces mediante el aprendizaje (Matich, 2001).

De lo explicado anteriormente es importante aclarar que, la función del diseñador de una red neuronal es únicamente la de la obtención de la arquitectura apropiada. No es problema del diseñador el cómo la red aprenderá a discriminar. Sin embargo, sí es necesario que desarrolle un buen algoritmo de aprendizaje que le proporcione a la red la capacidad de discriminar mediante un entrenamiento con patrones.

Autoorganización

Las redes neuronales emplean su capacidad de aprendizaje adaptativo para autoorganizar la información que reciben durante el aprendizaje y/o la operación. Mientras que el aprendizaje es la modificación de cada elemento procesal, la autoorganización consiste en la modificación de la red neuronal completa para llevar a cabo un objetivo en específico.

Cuando las redes neuronales se usan para reconocer ciertas clases de patrones, estas, autoorganizan la información usada.

Esta autoorganización es la que provoca la generalización: facultad de las redes neuronales de responder apropiadamente cuando se les presentan datos o situaciones a

las que no habían sido expuestas anteriormente (Hilera & Martínez, 1995). Esta característica es muy importante cuando se tienen que solucionar problemas en los cuales la información de entrada no es muy clara.

Tolerancia a Fallos

Comparadas con los sistemas computacionales tradicionales donde estos pierden su funcionalidad cuando sufren un error de memoria, las redes neuronales no lo hacen, si se produce un fallo en un número no muy grande de neuronas y a pesar de que el comportamiento del sistema se vea influenciado, no sufren una caída repentina.

Hay dos aspectos distintos respecto a la tolerancia a fallos:

- Las redes pueden aprender a reconocer patrones con ruido, distorsionados o incompletos. Esta es una tolerancia a fallos respecto a los datos.
- Las redes pueden seguir realizando su función, con cierta degradación, aunque se destruya parte de la red.

La razón por la que las redes neuronales son tolerantes a los fallos es que tienen su información distribuida en las conexiones entre neuronas, existiendo cierto grado de redundancia en este tipo de almacenamiento. Esto se logra gracias a que las redes neuronales almacenan la información de manera no localizada.

Operación en Tiempo Real

Una de las mayores prioridades, caso en la totalidad de las áreas de aplicación, es la necesidad de realizar procesos con datos de forma muy rápida. Las redes neuronales se adaptan bien a esto debido a su implementación paralela. Para que la mayoría de las redes puedan operar en un entorno de tiempo real, la necesidad de cambio en los pesos de las conexiones o entrenamiento en mínimo.

3.4 Topologías Principales de las Redes Neuronales

Se entiende como topología o arquitectura de una red neuronal a la organización y disposición de las neuronas, donde se pueden formar capas o asociaciones de neuronas que se encuentran más o menos alejadas de las entradas y salidas de la red. Por lo tanto,

se entienden como parámetros fundamentales de una red: el número de capas, cuántas neuronas hay en cada capa, su grado de conectividad y su tipo de conexión.

Es así como se pueden distinguir dos clases de redes neuronales según su número de capas.

- Redes monocapa
- Rede multicapa.

Redes monocapa

Para este tipo de redes, las conexiones laterales se establecen entre neuronas pertenecientes a la única capa que constituye a la red, lo anteriormente descrito se encuentra ejemplificado en la **Figura 3.4**. Generalmente son utilizadas para tareas relacionadas con auto-asociación, es decir, se regenera la información de entrada a la red cuando esta se presenta de forma incompleta o distorsionada (Hilera & Martínez, 1995).

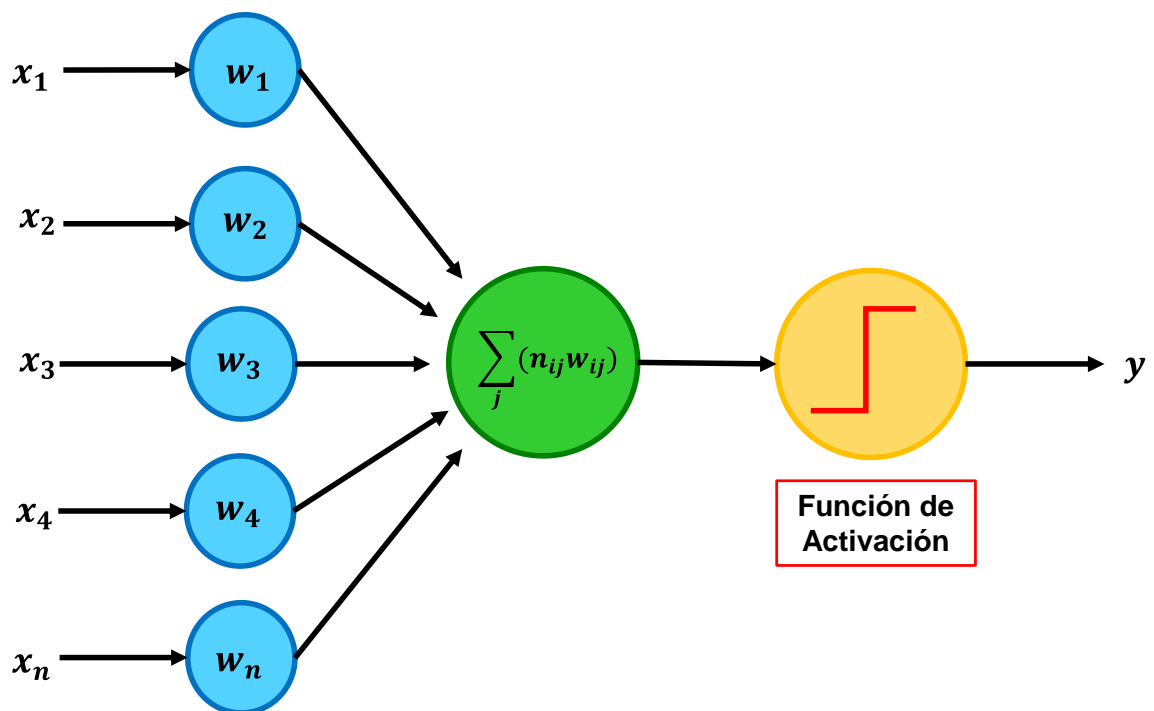


Figura 3.4. Red Neuronal Monocapa.

Redes multicapa

Las redes multicapa son aquellas que disponen de agrupaciones de neuronas en dos o más niveles o capas. Para estos casos, es posible distinguir a qué capa pertenece una neurona si se distingue qué señales recibe en la entrada y cuál es el destino de la señal de salida.

Cuando las neuronas de una entrada reciben señales de una capa posterior y envían señales a una capa de salida desde una capa anterior, se les puede llamar a dichas conexiones como *feedforward*. Por otra parte, para el caso de que las conexiones de salida de neuronas de capas posteriores se conecten a la entrada de capas anteriores se le conoce como conexiones *feedback*.

3.4.1 Función de Entrada

La neurona trata a todos los valores de entrada como si fueran uno solo, a esto se le conoce como entrada global. Para lograrlo, es necesario encontrar la manera de combinar las entradas de forma simple. Lo anterior se hace a través de la función de entrada, la cual es calculada a partir de un vector.

Es entonces cuando los valores de entrada se multiplican por los pesos anteriormente ingresados, dichos pesos al inicio del proceso de aprendizaje son designados al azar y modificados conforme la red aprende. Por lo tanto, los pesos cambian la influencia que tiene cada uno de los valores de entrada, es decir, que permiten que un gran valor de entrada tenga solamente una pequeña influencia, si estos son lo suficientemente pequeños.

A continuación, se presentan algunas de las funciones más utilizadas y conocidas.

- Suma ponderada de las entradas pesadas: se refiere a la suma de todos los valores de entrada a la neurona, multiplicándolos por su correspondiente peso.

$$\sum_j (n_{ij}w_{ij}), \rightarrow j = 1, 2, \dots, n. \quad \dots \dots \dots \quad (3.1)$$

- Producto de las entradas pesadas: es el producto de todos los valores de entrada a la neurona, multiplicados por sus correspondientes pesos.

$$\prod_j (n_{ij} w_{ij}), \rightarrow j = 1, 2, \dots, n. \dots\dots\dots (3.2)$$

- Máximo de las entradas pesadas: solamente se toma en consideración el valor de entrada más fuerte, previamente multiplicado por su correspondiente peso.

$$\text{Max} (n_{ij} w_{ij}), \rightarrow j = 1, 2, \dots, n. \dots\dots\dots (3.3)$$

3.4.2 Función de Activación

De igual forma que una neurona biológica que puede estar activa o inactiva, las neuronas artificiales también tienen un estado de activación, algunas de estas solamente tiene dos similarmente que las biológicas, pero otras pueden tomar cualquier valor dentro de un cierto rango.

La función de activación calcula el estado de actividad de una neurona; transformando la entrada global (menos el umbral, θ_i) en un valor (estado) de activación, cuyo rango normalmente va de (0,1) o de (-1,1). Esto es así ya que una neurona puede estar o totalmente activa o completamente inactiva.

Las funciones de activación más comunes se presentan a continuación.

- Función lineal:
 Por encima o por debajo de esta zona se fija la salida en 1 o -1. Cuando $a = 1$, la salida es igual que la entrada **Figura 3.5**.

$$f(x) = \begin{cases} -1, & x \leq -1/a \\ a * x, & -\frac{1}{a} < x < \frac{1}{a}, \rightarrow x = \text{gin} - \theta_i \text{ y } a > 0 \\ 1, & x \geq 1/a \end{cases} \dots\dots\dots (3.4)$$

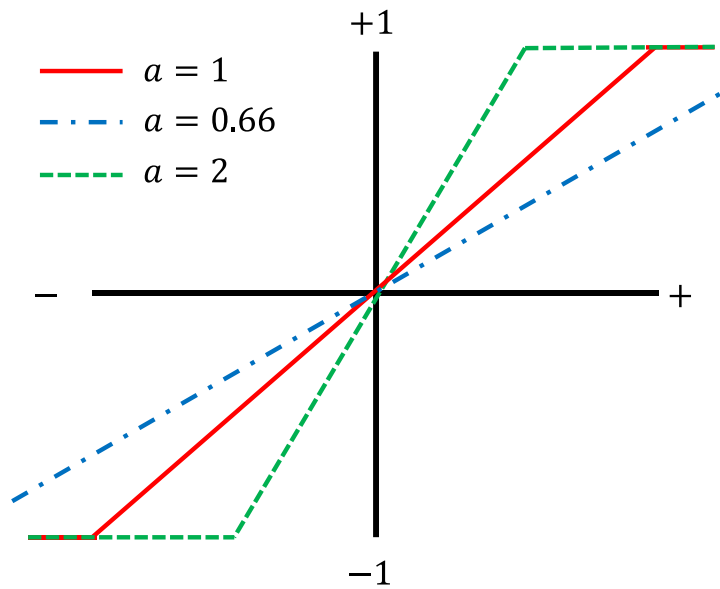


Figura 3.5. Función Lineal de Activación. Modificada de (Match, 2001).

- Función sigmoide:

Los valores de la salida que proporciona esta función están comprendidos en un rango entre 0 y 1. Al modificar el valor de g , se ve afectada la pendiente de la función de activación **Figura 3.6**.

$$f(x) = \frac{1}{1 + e^{-gx}}, \quad \rightarrow \quad x = gin_i - \theta_i. \quad \dots\dots\dots (3.5)$$

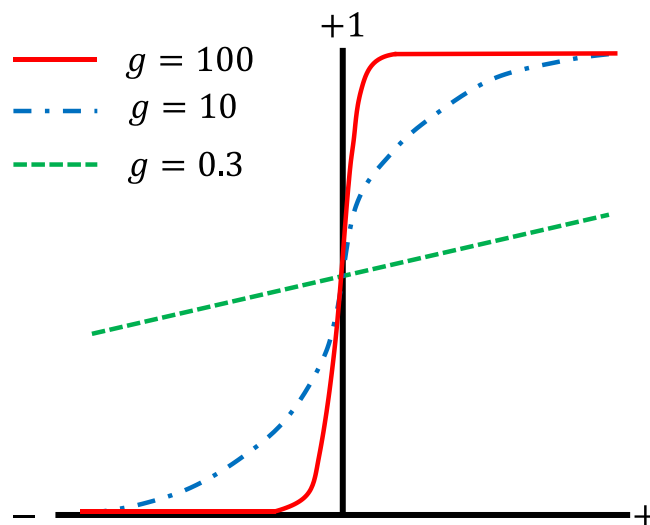


Figura 3.6. Función de Activación Sigmoidea. Modificada de (Match, 2001).

- Función tangente hiperbólica:

Los valores de salida de la función tangente hiperbólica están comprendidos dentro del rango de -1 a 1. Al modificar el valor de g , se ve afectada la pendiente de la función de activación **Figura 3.7**.

$$f(x) = \frac{e^{gx} - e^{-gx}}{e^{gx} + e^{-gx}}, \quad \rightarrow \quad x = \text{fin}_i - \theta_i. \quad \dots\dots\dots (3.6)$$

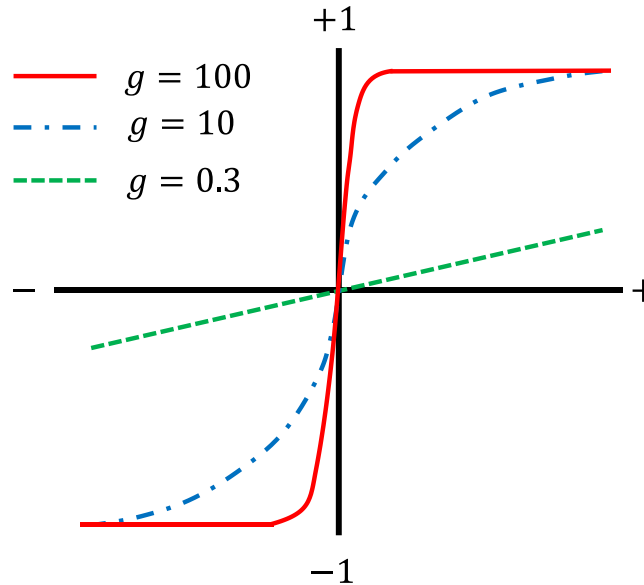


Figura 3.7. Función de Activación Hiperbólica. Modificado de (Matich, 2001).

3.4.3 Función de Salida

Es el último componente que una neurona necesita. El valor resultante de esta función es la salida de la neurona i (out_i), por lo tanto, la función de salida determina que valor se transfiere a las neuronas vinculadas. Si la función de activación está por debajo de un umbral determinado, ninguna salida se pasa a la neurona subsiguiente.

Las funciones de salida más comunes son:

- Ninguna: este es el tipo de función más sencillo, tal que la salida es la misma que la entrada. Es también llamada función identidad.

- Binaria:

$$B \begin{cases} 1 & \text{si } act \geq \xi_i \\ 0 & \text{de lo contrario} \end{cases}, \quad \xi_i \text{ es el umbral.} \quad \dots \quad (3.7)$$

Conexiones Entre Neuronas

Las conexiones que unen las neuronas tienen asociado un peso que es el que hace que la red adquiera conocimiento. Una neurona recibe un conjunto de señales que le dan la información del estado de activación de todas las neuronas con las que se encuentra conectada. Cada conexión de la neurona i con la neurona j está ponderada por un peso w_{ij} . Normalmente, como una forma para simplificar, se considera que el efecto de cada señal es aditivo, de tal forma que la entrada neta que recibe una neurona net_j es la suma del producto de cada señal individual por el valor de la sinapsis que conecta ambas neuronas.

3.4.4 Mecanismos de Aprendizaje

Se le conoce como proceso de entrenamiento al procedimiento donde una red neuronal aprende a calcular la salida correcta a partir de los datos de entrada del conjunto de ejemplo, el cual es llamado conjunto de entrenamiento. Por lo tanto, se puede definir al aprendizaje como el proceso por el cual una red neuronal modifica sus pesos como respuesta a una información de entrada, y que se ve reflejado en la modificación, destrucción o creación de conexiones entre neuronas. Entendiendo como destrucción cuando el valor del peso de una red neuronal es igual a cero, de forma contraria, se habla de una creación de conexión cuando el peso toma un valor distinto de cero.

Se puede decir que el aprendizaje para una red neuronal ha terminado cuando las tendencias de cambio en los pesos de las conexiones a lo largo del tiempo son igual a cero, por lo tanto, se puede concluir que la red ha aprendido.

Existen dos métodos de aprendizaje de gran importancia:

- Aprendizaje no Supervisado.
- Aprendizaje Supervisado.

Aprendizaje No Supervisado

También conocidas como redes con aprendizaje autosupervisado o autoorganizado; este tipo de red no requiere de un agente externo para realizar el ajuste de sus pesos sinápticos. La red no recibe información alguna que pueda indicar si la salida es correcta o no, por ello, se dice que estos tipos de redes son capaces de autoorganizarse.

En cuanto a los procesos de aprendizaje de estas redes se pueden clasificar en dos tipos: aprendizaje Hebbiano y aprendizaje competitivo y cooperativo.

Aprendizaje Supervisado

Este tipo de aprendizaje se caracteriza por la realización de un entrenamiento de la red controlado por un agente externo o supervisor, el cual determinará la respuesta considerada correcta y que la red deberá obtener a partir de una entrada también controlada. En caso de no coincidir la salida con lo que se busca, se procede a realizar la modificación de los pesos de las conexiones con la finalidad de obtener la salida deseada o bien una salida aproximada a esta.

A su vez este tipo de aprendizaje puede clasificarse en tres formas según cómo se lleve a cabo:

- **Aprendizaje por Corrección de Error:** Se lleva a cabo ajustando los pesos de las conexiones de la red en función de la diferencia entre los valores deseados y los obtenidos en la salida, de acuerdo con el margen de error cometido.
- **Aprendizaje por Esfuerzo:** Este tipo de aprendizaje es más lento que el anterior ya que la salida de la red no puede compararse al valor deseado para poder obtener el error cometido. El funcionamiento de esta red es similar a los métodos numéricos. La función del supervisor en este caso es indicar mediante una señal de refuerzo si la salida de la red se ajusta a la deseada (*éxito* = +1 o *fracaso* = -1) y en función de ello se ajustan los pesos basándose en un mecanismo de probabilidades.
- **Aprendizaje Estocástico:** Básicamente consiste en realizar cambios aleatorios en los pesos y evaluar su efecto a partir del objetivo deseado y de atribuciones de probabilidad.

Algoritmo de Aprendizaje: Backpropagation

A continuación, se explica el método de aprendizaje supervisado conocido como *backpropagation*, uno de los métodos más usados en la solución del tipo de problemas como el planteado en este trabajo.

Minsky y Papert en 1969 demostraron que una red de una sola capa feedforward puede tener muchas restricciones. Dichas limitaciones fueron un factor significativo para que declinara el interés para seguir desarrollando modelos de redes neuronales en los años 70. El descubrimiento de un método general efectivo de entrenamiento de una red neuronal multicapa, *backpropagation* fue propuesto en 1986 por Rumelhar, Hinton y Williams. La idea central de esta solución es que los errores de las unidades de las capas ocultas son determinadas por retro-propagación de errores desde las unidades de la capa de salida.

La propagación hacia atrás de errores o retro-propagación (del inglés *backpropagation*) es un método de cálculo del gradiente utilizado en algoritmos de aprendizaje supervisado para entrenar redes neuronales artificiales. El método emplea un ciclo propagación – adaptación de dos fases. Una vez que se ha aplicado un patrón a la entrada de la red como estímulo, éste se propaga desde la primera capa a través de las capas siguientes de la red, hasta generar una salida. La señal de salida se compara con la salida deseada y se calcula una señal de error para cada una de las salidas.

Las salidas de error se propagan hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la capa oculta que contribuyen directamente a la salida. Sin embargo, las neuronas de la capa oculta solo reciben una fracción de la señal total del error, basándose aproximadamente en la contribución relativa que haya aportado cada neurona a la salida original. Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido una señal de error que describa su contribución relativa al error total.

La importancia de este proceso consiste en que, a medida que se entrena la red, las neuronas de las capas intermedias se organizan a sí mismas de tal modo que las distintas neuronas aprenden a reconocer distintas características del espacio total de entrada. Después del entrenamiento, cuando se les presente un patrón arbitrario de entrada que

contenga ruido o que esté incompleto, las neuronas de la capa oculta de la red responderán con una salida activa si la nueva entrada contiene un patrón que se asemeje a aquella característica que las neuronas individuales hayan aprendido a reconocer durante su entrenamiento.

El algoritmo *backpropagation* se encarga de buscar el valor mínimo del error para los pesos, esto se logra mediante una técnica conocida como regla del gradiente descendiente o regla delta. Los pasos de este algoritmo de manera generalizada son los siguiente:

- Cálculo del error, es decir, qué tanto se alejan los valores de salida de la red de aquellos valores deseados o considerados correctos.
- Error mínimo, lo que se busca en este paso es determinar si el error obtenido es el mínimo posible.
- Actualizar los parámetros, si el error calculado anteriormente es muy grande, es necesario actualizar los valores de los pesos del modelo. El proceso anterior se repite hasta que el error calculado sea el mínimo posible.

Modelo listo, una vez se ha encontrado el error mínimo, es posible obtener los valores de salida deseados predichos a partir de una entrada.

La **Figura 3.8** muestra, en diagrama de bloques el proceso anteriormente descrito.

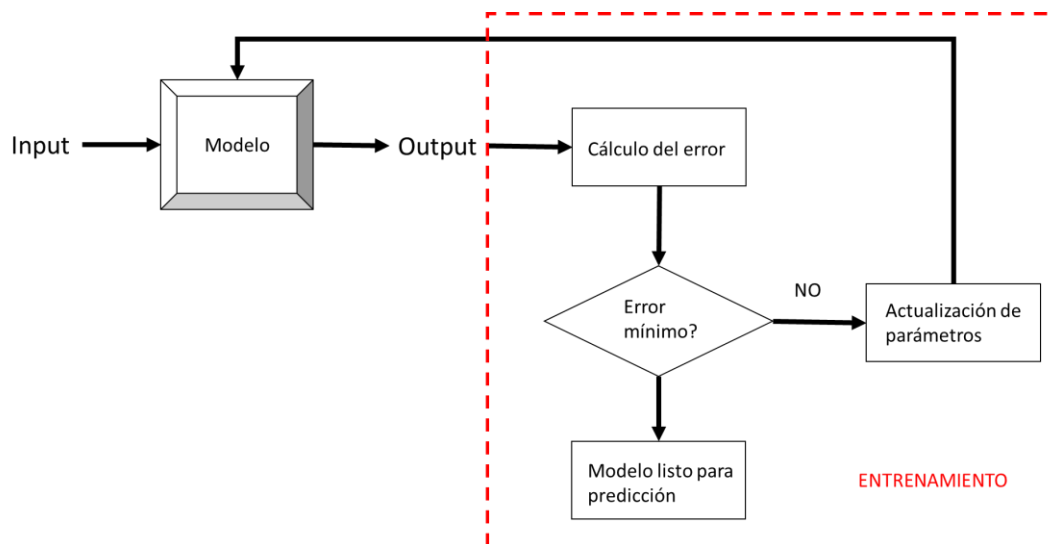


Figura 3.8. Esquema de Bloques del Aprendizaje Backpropagation

La naturaleza general del método backpropagation (multicapa, red feedforward entrenado por backpropagation) hace que éste pueda ser usado para resolver problemas en distintas áreas del conocimiento. El entrenamiento de una red de retro-propagación como la que se esquematiza en la **Figura 3.9**, se puede dividir en las siguientes tres etapas.

- Entrenamiento Feedforward con señales de entrada: Cada nodo de entrada recibe una señal, la cual es enviada a todas las unidades ocultas. Cada unidad oculta determina una activación para enviar señales a otros nodos ocultos hasta llegar a todos los nodos de las capas de salida.
- Retro-propagación de errores: Cada nodo de salida compara su valor de activación o calculado, con el valor de salida deseada. Basado en estas diferencias, el error es propagado hacia los nodos anteriores.
- Ajuste de pesos: Todos los pesos se calculan simultáneamente basados en los errores que fueron propagados en la red.

Este método puede ser considerado como la generalización de la regla delta para la función de activación en redes de múltiples capas. La regla delta solo trabaja para la capa de salida, backpropagation o generalización de la regla delta es un camino para crear valores deseados en las capas ocultas.

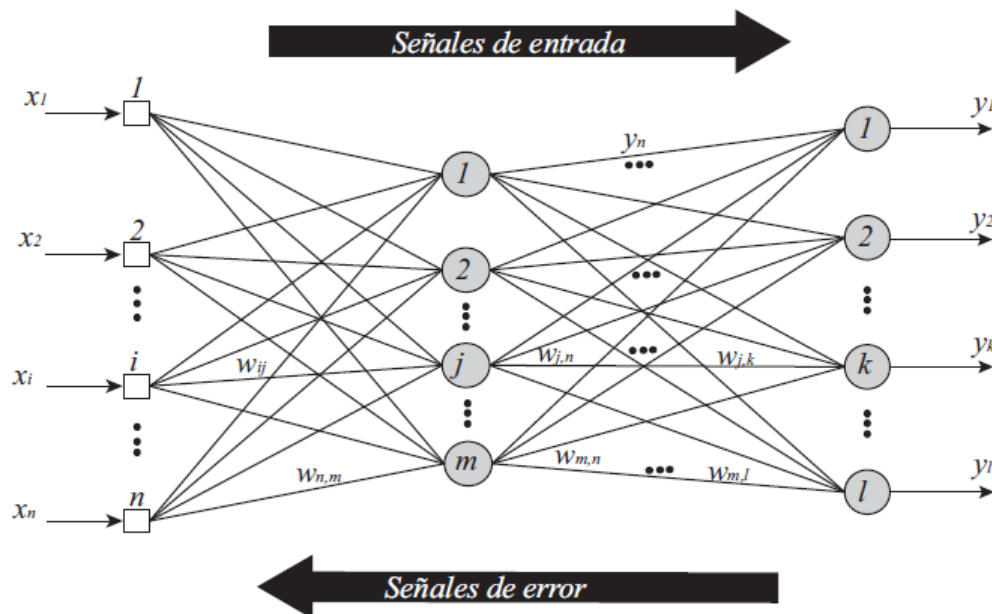


Figura 3.9. Esquemización del algoritmo backpropagation. Tomado de (Hernández Ambrosio, 2017).

Etapas del entrenamiento

Las etapas por seguir para ejecutar el algoritmo de retro-propagación de forma correcta son las enunciadas a continuación.

Etapas 0: Se inicializan los pesos como un conjunto de valores aleatorios.

Etapas 1: Mientras la condición “parar” sea falsa, se deben realizar los pasos 2 a 9 aquí nombrados.

Etapas 2: Para cada par de entrenamiento, ejecutar las etapas 3 a 8.

Feedforward

Etapas 3: Para cada unidad de entrada ($x_i, \rightarrow i = 1, 2, 3, \dots, n$), recibe una señal x_i y se envía esta señal a todas las unidades de la capa siguiente (unidades de la capa oculta).

Etapas 4: Cada unidad oculta ($a_{inj}, \rightarrow j = 1, 2, 3, \dots, m$) suma los pesos de las señales de entrada o realiza la operación correspondiente según la función de entrada escogida.

$$a_{inj} = net_{inj} = \sum_{i=1}^n x_i w_{ij}, \dots\dots\dots (3.8)$$

se aplica la función de activación para el cálculo de la señal de salida como:

$$a_j = f(net_{inj}), \dots\dots\dots (3.9)$$

y se envía la señal a todas las unidades de la capa siguiente.

Etapas 5: Suma de cada unidad de salida ($y_k, \rightarrow k = 1, 2, 3, \dots, l$) con los pesos en la señal de entrada,

$$y_{ink} = net_k = \sum_{j=1}^m a_j w_{j,k}, \dots\dots\dots (3.10)$$

aplicando la función de activación se obtiene la señal de salida,

$$y_k = f(y_{ink}). \quad \dots\dots\dots (3.11)$$

Retro-propagación del error

Etapa 6: Cada unidad de salida ($y_k = 1, 2, 3, \dots, l$) recibe el valor deseado correspondiente al patrón de entrenamiento, para el cálculo del error:

$$\delta_k = (z_k - a_k)f'(y_{ink}), \quad \dots\dots\dots (3.12)$$

con este valor se puede calcular el término de corrección de peso (usado para actualizar el peso $w_{j,k}$).

Etapa 7: Cada unidad oculta ($a_j, \rightarrow j = 1, 2, 3, \dots, m$) suma a la entrada delta,

$$\delta_{inj} = \sum_{k=1}^l \delta_k w_{j,k}, \quad \dots\dots\dots (3.13)$$

multiplicando por la derivada de la función de activación para calcular el error, para las j -ésimas unidades ocultas,

$$\delta_j = \delta_{inj}f'(a_{inj}). \quad \dots\dots\dots (3.14)$$

Actualización de Pesos

Etapa 8: Para cada unidad de salida ($y_k, \rightarrow k = 1, 2, 3, \dots, l$) se deberá actualizar su correspondiente peso. La regla de cambio de pesos para la capa de salida es la siguiente:

$$w_{j,k}(Nuevo) = w_{j,k}(Anterior) + \Delta w_{j,k}, \quad \dots\dots\dots (3.15)$$

donde

$$\Delta w_{j,k} = \eta \delta_k a_j.$$

Ahora, para realizar la actualización de los pesos en las unidades de entrada a una unidad oculta, se procede a hacer el siguiente cálculo:

$$w_{j,k}(\text{Nuevo}) = w_{j,k}(\text{Anterior}) + \Delta w_{i,j}, \quad \dots \quad (3.16)$$

donde

$$\Delta w_{i,j} = \eta \delta_j a_j.$$

Etapas 9: Establecer la condición “detener”, cuando se cumpla lo establecido.

Capítulo 4: Resultados

En este capítulo se presentan las zonas de estudio, abordando sus características geológicas, geográficas y las relacionadas con el ámbito petrolero, así como las formaciones importantes de donde se tomaron las muestras para los análisis de núcleos y los registros de pozo.

Se muestran, además, la metodología que explota herramientas de ciencia de datos e inteligencia artificial, la comparación con resultados obtenidos mediante técnicas convencionales de determinación o estimación de la permeabilidad y se analizan las ventajas y desventajas de la propuesta.

4.1 Sobre la Kansas Geological Survey (KGS)

La KGS es la división de investigación y servicios de la Universidad de Kansas que investiga y provee información sobre los recursos geológicos, petroleros y de aguas subterráneas del estado de Kansas.

En esta división se persigue la investigación relacionada a temas de geología subsuperficial, recursos energéticos, agua subterránea y peligros medioambientales. Busca desarrollar herramientas y técnicas innovadoras, monitorear terremotos y el nivel de las aguas subterráneas, investigar asuntos relacionados con la calidad del agua, así como mapear la geología superficial del Estado. Todo descubrimiento, análisis y datos obtenidos son compartidos con la comunidad científica y el público en general a través de publicaciones, recursos online y presentaciones. La KGS almacena, además, miles de registros de pozos de agua, aceite y gas, que han sido llenados con el estado a través de varias décadas, así como miles de núcleos de roca traídos a la superficie de la tierra durante la perforación de pozos de aceite y gas.

De esto último descrito que se haya utilizado el repositorio en línea (KGS, 2020), así como las herramientas proporcionadas por la KGS para la obtención de datos y el desarrollo de este trabajo.

Repositorio en Línea

De manera concreta se explica a continuación el funcionamiento del repositorio en línea de datos de pozos de agua, petróleo y gas pertenecientes a la KGS. Esta explicación se encuentra resumida en un diagrama de bloques encontrado en el **Apéndice 1**.

Primeramente, es importante aclarar que este repositorio es completamente gratuito para el público en general, por lo tanto, no es necesario contar con ninguna clase de perfil o registro para acceder y revisar la información que la KGS proporciona.

Aclarado lo anterior, la página perteneciente a la KGS donde se encuentra el repositorio y se conjunta la información, puede clasificarse en dos categorías. En la primera puede encontrarse todo lo referente a publicaciones, artículos, educación y la visión y misión de la KGS. La segunda parte corresponde a los geodatos disponibles, información de las distintas secciones que tiene esta categoría clasificadas en: agua, energía, geofísica y geología. Cada una con sus particulares motores de búsqueda de información.

La sección a prestar atención es la de “energía” ya que en esta se encuentra información sobre los pozos de gas y aceite de Kansas, así como la producción de todos los campos en su territorio. Es posible acceder a mapas interactivos, publicaciones, cursos y tutoriales.

La información general necesaria para usar cualquiera de los buscadores disponibles es la siguiente:

- **County:** En caso de conocer en cuál de los 91 condados de disponibles se encuentra el pozo o el campo de interés, es de resaltar que en esta sección se puede acceder a todos los pozos de Kansas buscando con la opción “Cualquier condado” y no es necesario contar con otra información. Esta división puede observarse en la **Figura 4.1**, donde se aprecia en colores los campos petroleros del estado. El condado con el recuadro rojo se resalta ya que será usado para la ampliación de las otras imágenes útiles para explicar la forma de referenciar las áreas donde se encuentran los pozos y parte de los campos.
- **Wells:** Se refiere al nombre del pozo buscado.

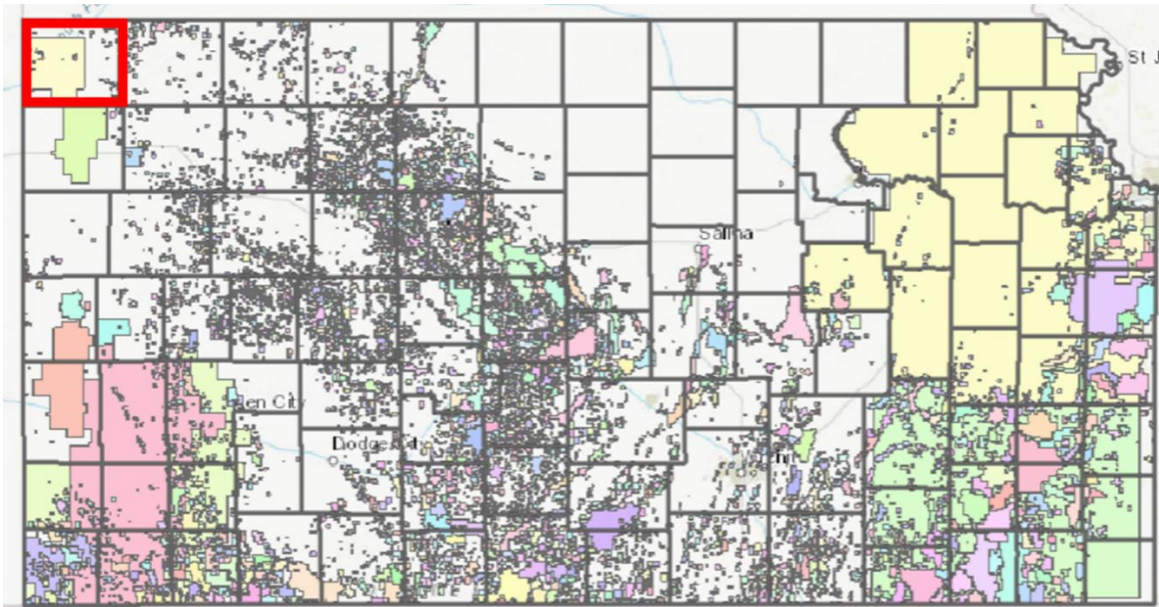


Figura 4.1. División del estado de Kansas en Condados. Tomado de la KGS

En cuanto a la información particular en los buscadores se tienen las siguientes opciones:

- **Township:** Se refiere al municipio en la que se encuentra el pozo, esta división es posible apreciarla en la **Figura 4.2**.
- **Range:** Subdivisión que junto con *township* sirve para ubicar en los mapas interactivos campos y pozos, apreciable igualmente en la figura 4.2.

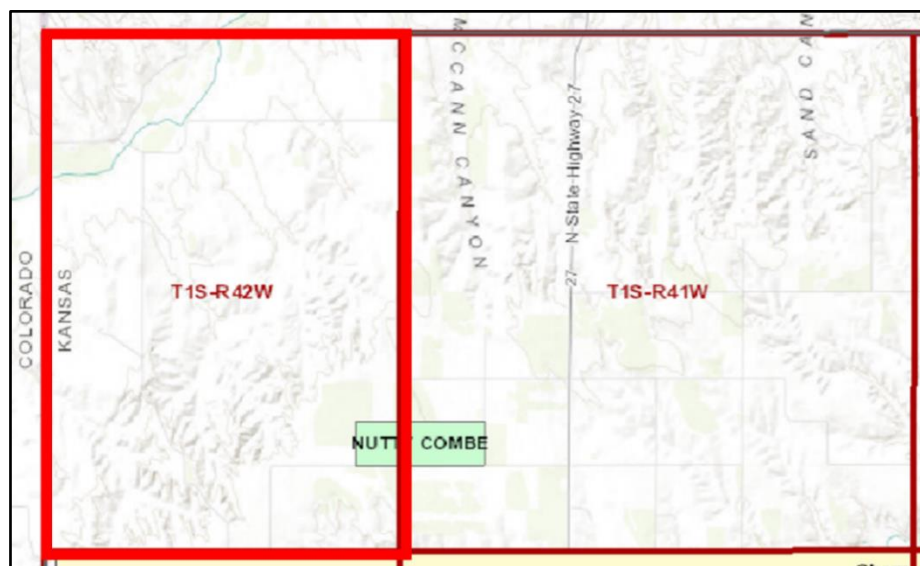


Figura 4.2. División en municipio (*Township*) y rango (*Range*) de cada uno de los condados. Tomado de la KGS

- **Section:** Sección de una cuadrícula donde se subdivide el área dada por la conjunción de *township* y *Range*, con la finalidad de ubicar de forma más particular y precisa algún pozo o sección de un campo, visible en la **Figura 4.3**.

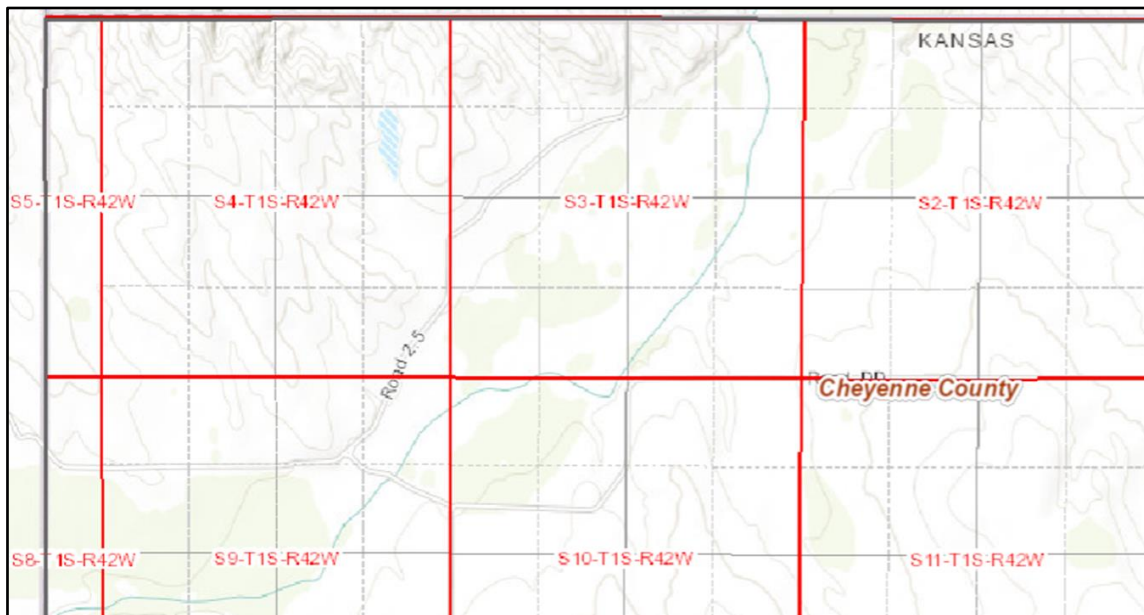


Figura 4.3. División del área en secciones. Tomado de la KGS

- **Lease:** Hace referencia al nombre del arrendatario del área donde se llevan a cabo las operaciones de extracción o perforación de los pozos.
- **Operator:** Se refiere al nombre del operador del campo donde se encuentra el pozo buscado.
- **API well:** Es un número asignado a cada uno de los pozos.

Finalmente, contando con cualquiera de los dos tipos de información es posible acceder a los buscadores de la sección de pozos de aceite y gas que llevan por nombre:

- **Lista maestra de pozos de gas y aceite:** donde se despliega toda la información básica por pozo de todos los pozos en Kansas.
- **Encabezados de registros geofísicos de pozo:** aquí es posible visualizar los encabezados de todos los pozos que cuenten con registros, cuenta con

información como ubicación geográfica, registros tomados, rango de profundidad a la que se realizaron los registros, entre otros.

- **Archivos .LAS disponibles:** se despliegan todos los pozos que contengan archivos .LAS, que son aquellos que precisan toda la información de los registros tomados por pozo.
- **Análisis de núcleos:** este último buscador contiene todo lo referente a resultados de análisis de núcleos realizados por pozo.

La información que se puede obtener y descargar de estos cuatro buscadores es la siguiente:

- Columna litológica.
- Carriles de registros escaneados.
- Reportes de los estudios de núcleos en laboratorio.
- Datos de Registros.
- Archivos .LAS con la información de los registros.

Finalmente, es posible tener acceso tanto desde el menú como desde los pozos desplegados en los buscadores hacia un mapa interactivo donde se muestran todos los pozos, así como los campos en Kansas. Este mapa interactivo mostrado en la **Figura 4.4** cuenta con herramientas de búsqueda y filtrado de pozos.

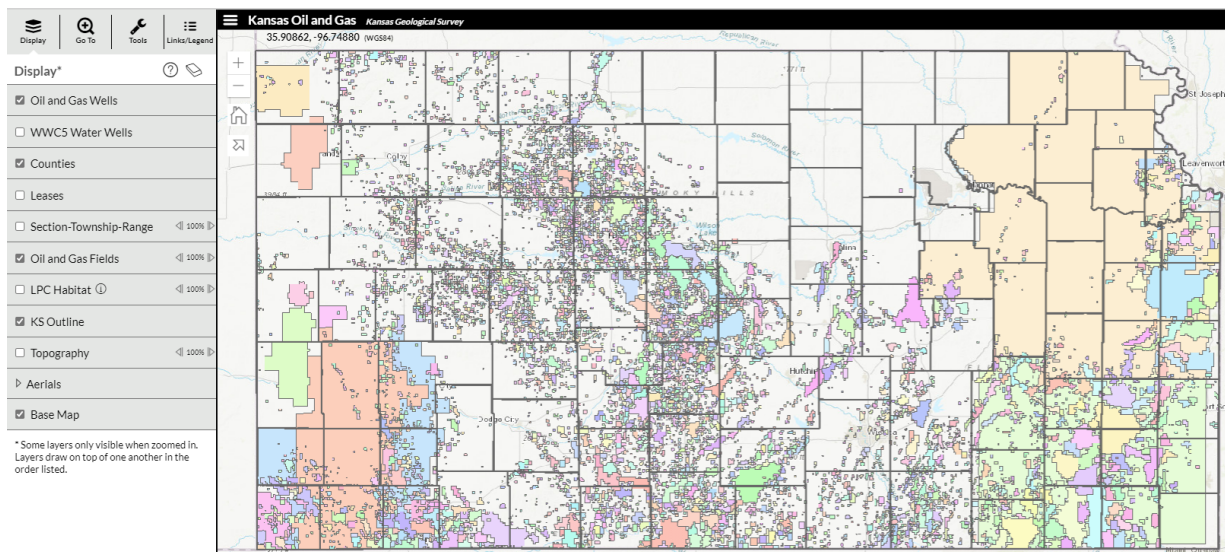


Figura 4.4. Mapa Interactivo de la KGS.

4.2 Resumen de Casos

Las áreas de estudio se encuentran en la zona sur-central, suroeste y noreste del estado de Kansas, Estados Unidos. De aquí se obtuvo información de siete pozos que contaban con los datos de distintos registros y además tenían los resultados de los análisis de núcleos en laboratorio. La información que se obtuvo de estos pozos es la siguiente:

- un registro de inducción (ILD) con unidades en ohm-m,
- un registro de rayos gamma (GR) con unidades GAPI,
- un registro de porosidad neutrón (NPHI) con unidades adimensionales (fracción),

se cuenta, además, con

- la permeabilidad absoluta al líquido en milidarcys obtenida mediante estudios de núcleo en laboratorio,
- la porosidad medida de la muestra,
- la densidad de grano, y
- las saturaciones en porcentaje del volumen poroso con las que se calculó la porosidad.

Un pozo (con faltantes en permeabilidades de la columna muestreada) servirá como una de las formas de validación de los modelos trabajados.

A continuación, se detallan las características litológicas y estratigráficas de las formaciones productoras a las que pertenecen los pozos. El conjunto está conformado por seis pozos productores de aceite y uno productor de gas ubicados geográficamente según los estatutos de la Kansas Geological Survey que se explicaron en la sección anterior (*Township, Range, Section*) como sigue:

- Seis de estos pozos, cinco productores de aceite y uno de gas, se encuentran identificados dentro de los condados de Morton y Stanton con los marcadores S19-T31S-R40W, S23-T29S-R41W, S25-T31S-R41W y S3-T30S-R25W (S: sección, T: municipio, R: rango).
Cuatro de ellos distribuidos en un área de aproximadamente 16.74 km² (**Figura 4.5**).

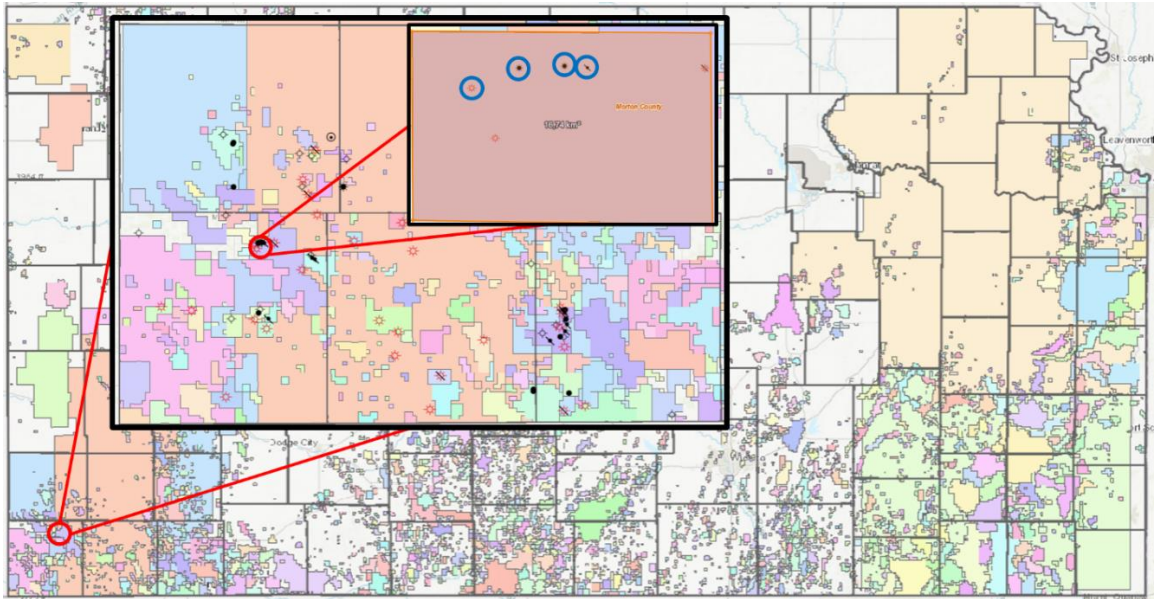


Figura 4.5. Localización de los pozos en los condados de Morton y Stanton. Tomada de KGS

El quinto pozo se encuentra en el condado de Stanton 20 km al norte de la aglomeración de los cuatro pozos mencionados anteriormente. Es importante resaltar que todos estos pozos pertenecen al campo Hugoton Gas remarcado en amarillo en la **Figura 4.6**.

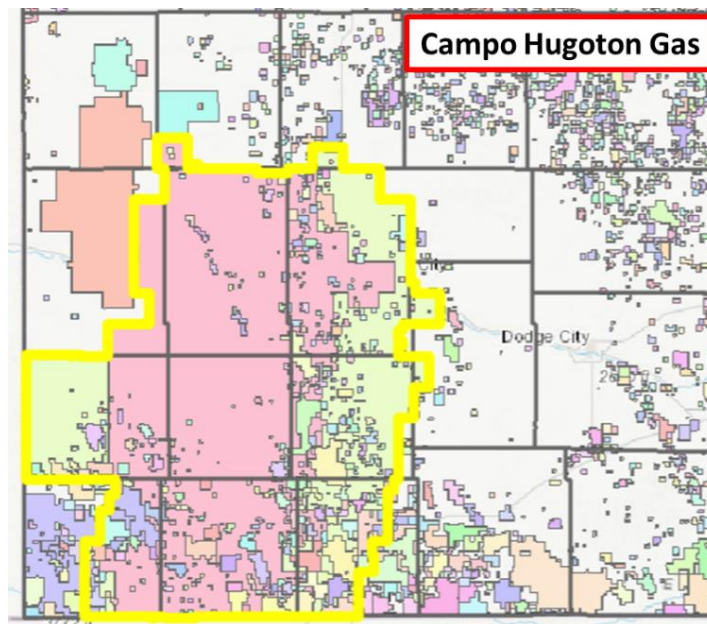


Figura 4.6. Área del Campo Hugoton Gas. Tomado de KGS

El sexto está a 140 km al este de la agrupación del condado de Morton en el campo Norcan East (**Figura 4.7**).

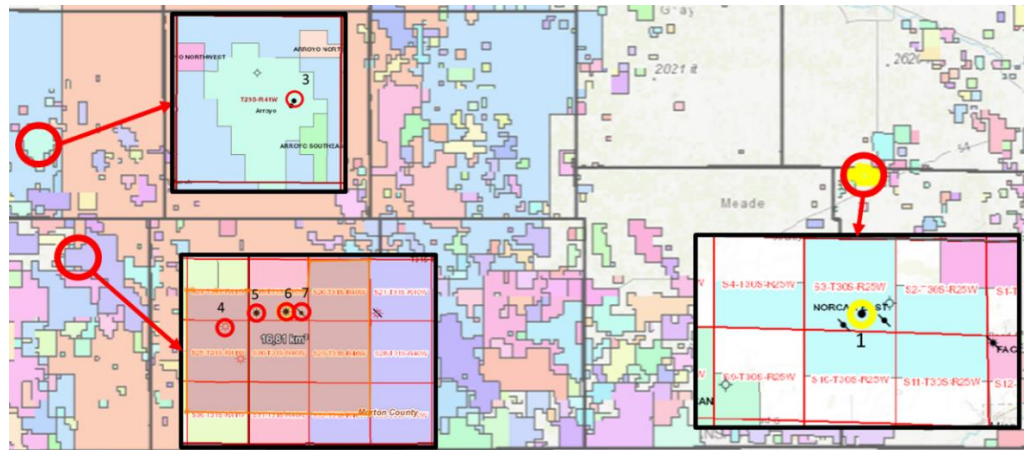


Figura 4.7. Distribución de los Pozos en el Estado de Kansas.

Las tres zonas que agrupan a seis de los siete pozos antes mencionados se encuentran en campos productores pertenecientes a los grupos geológicos Morrow y Atokan del Pennsilvánico inferior. Estas rocas se distribuyen en el este de Kansas en el embalse de Hugoton (**Figura 4.8**), se encuentran cubiertas por las formaciones pertenecientes al Pennsilvánico medio (calizas y lutitas) y además por encima de las rocas típicas del Mississippico temprano (calizas y areniscas).

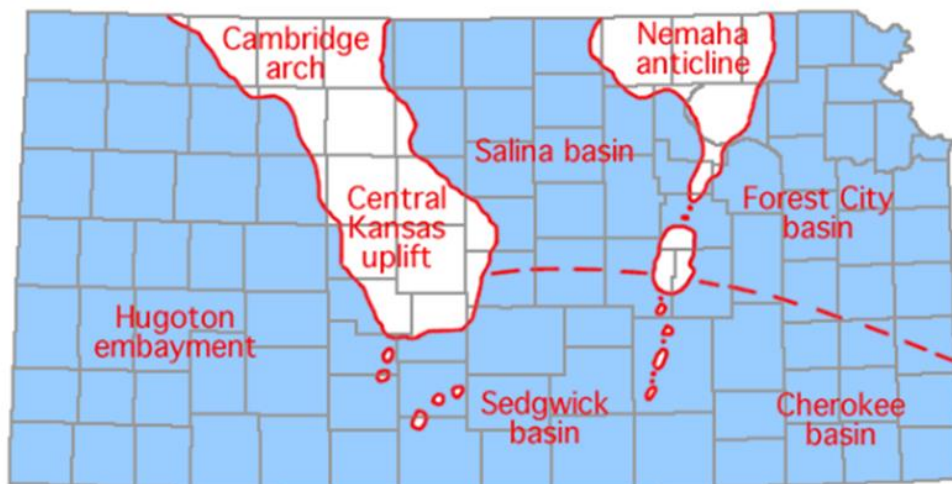


Figura 4.8. Provincias Estructurales y Distribución de Rocas del Mississippico y el Pennsilvánico en Kansas. Tomado de (Goebel, 1966).

Las rocas de la etapa Morrowan están compuestas de 600 ft (183 m) de lutitas, calizas y areniscas, siendo estas últimas especialmente abundantes en las formaciones del grupo Morrow que producen principalmente gas desde el nivel suroeste de Kansas. La producción de aceite se extiende hacia el norte desde esta zona para formar un área triangular de producción con una cúspide en el norte hacia Wallace County. El intervalo Atoka no es productivo en Kansas, pero lo es en la zona sur de Oklahoma y Texas, esto es apreciable en la **Figura 4.9**.

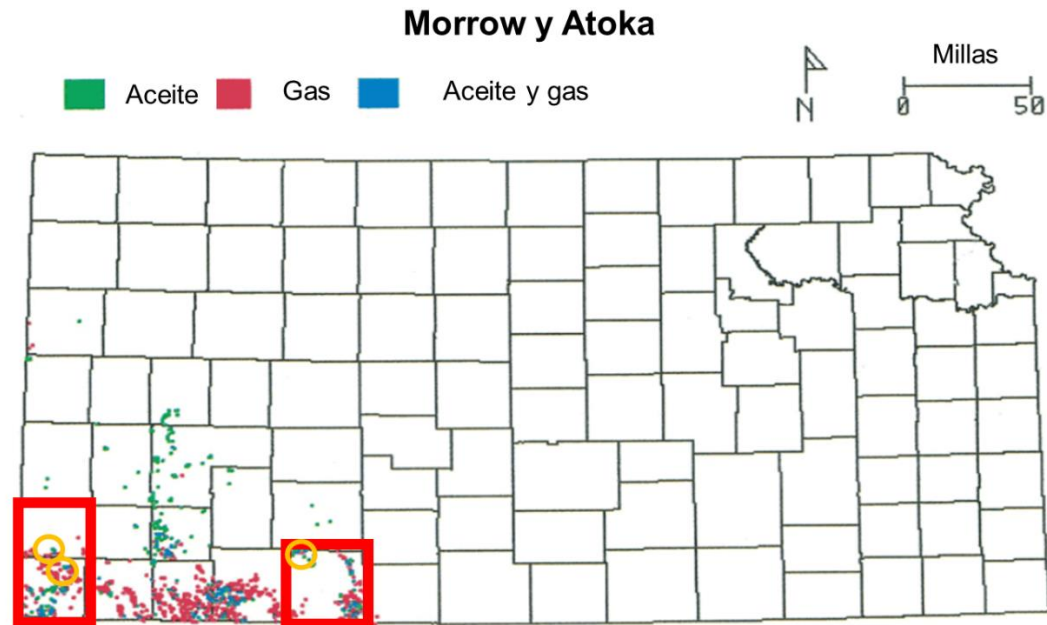


Figura 4.9. Campos de gas del sistema Morrow y Atoka en Kansas. Modificado de KGS.

Los sedimentos de los grupos geológicos de Morrow y Atoka fueron depositados en un gran embalse que se extiende hacia el norte desde la cuenca de Anadarki situada en Texas y el oeste de Oklahoma. El embalse cubre la mayor parte este de Colorado y oeste de Kansas, donde estos sedimentos se abrieron paso hacia el este y norte a lo largo de la zona que se extiende desde el condado de Cheyenne en el noroeste de Kansas hasta los condados de Comanche y Clark en la zona sur-centro de Kansas. El grosor máximo del intervalo productor en Kansas supera los 500 pies (150m) (Rascoe & Adler, 1983), lo que se aprecia en la **Figura 4.10** encerrado en los recuadros rojos.

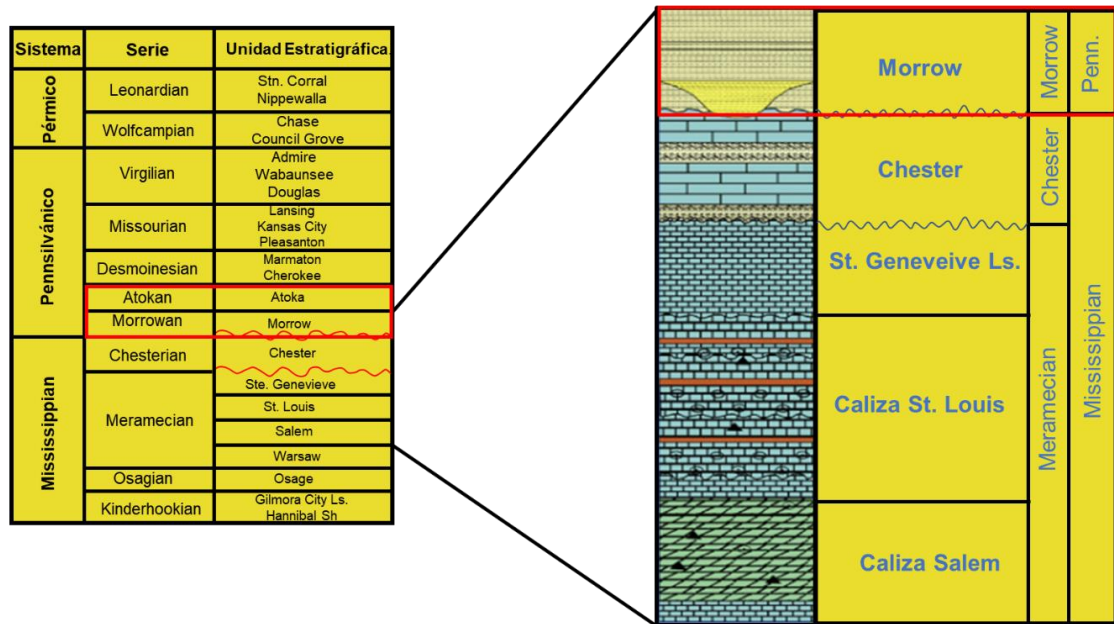


Figura 4.10. Columna Estratigráfica del Suroeste de Kansas (Morrow-Atoka). Modificado de KGS.

La formación Morrow ha sido dividida en dos unidades, la unidad inferior y la unidad superior. La unidad Morrow inferior tradicionalmente ha sido interpretada como lutitas costa fuera y areniscas de la línea de costa, mientras que la unidad superior como lutitas marinas que fueron encerradas por secuencias transgresivas. Depósitos marinos marginales y someros de la arenisca de Morrow se acumularon en el embalse de Hugoton (Wheeler, et al., 1990). Existe, además, una sección conocida como Morrow medio o Caliza “Squaw Belly” (Puckette et al., 1996) la cual es una unidad de caliza con presencia de lutitas que separa la sección arenosa Morrow inferior de la superior.

Playas de islas barrera y arenas marinas costa fuera han sido descritas en la formación Morrow inferior (McManus, 1959, Adams, 1964) que son comúnmente referidas como “Key Sandstones” (Rascoe & Adler, 1983). Las rocas que tienen función de yacimiento son lenticulares que van de un rango desde pobremente hasta bien clasificadas, con un tamaño de grano variando desde muy fino hasta grueso, comúnmente con poros parcialmente llenos de calcita, dolomita, cuarzo y caolinita. Las condiciones deposicionales del estrato Morrow superior se vieron dominadas por depósitos fluviales-deltaicos que reflejan condiciones menores de

una regresión marina (**Figura 4.11**). Estas areniscas son normalmente de grano grande, conglomeradas localmente donde además existe presencia de carbonatos y minerales de arcilla.

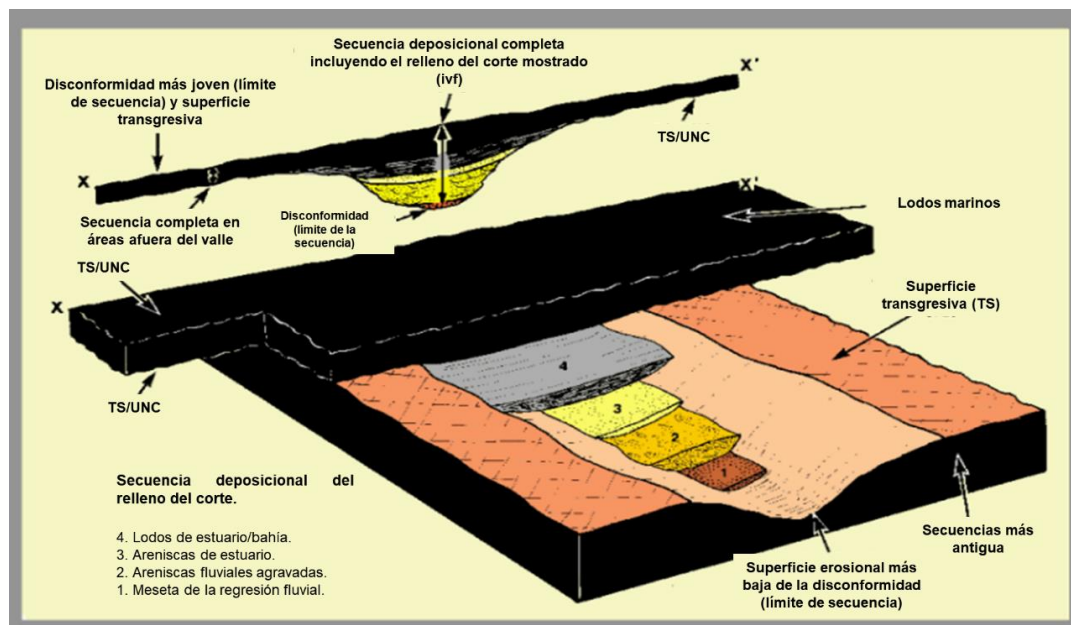


Figura 4.11. Condiciones Deposicionales de la Formación Morrow. Modificado de KGS

Una parte importante de los campos productores de aceite y gas en Kansas producen principalmente de areniscas lenticulares pertenecientes a la formación Morrow superior con grosores que van desde dos hasta 60 pies (0.6-18.2m). Trampas estratigráficas estructurales dominan los campos que del este de Kansas.

- Finalmente, el pozo productor de aceite con identificador de la zona S7-T4S-R14E pertenece al grupo Viola y Maquoketa del Ordovícico medio y superior. Este grupo geológico se encuentra esparcido en la zona noreste y norte central de Kansas aproximadamente en la misma distribución que la formación Simpson. Predomina la producción de aceite, pero existen zonas productoras de gas y aceite. Los yacimientos de las formaciones pertenecientes al grupo Viola predominan en esta área ya que Maquoketa no es una formación viable para la explotación **Figura 4.12**.

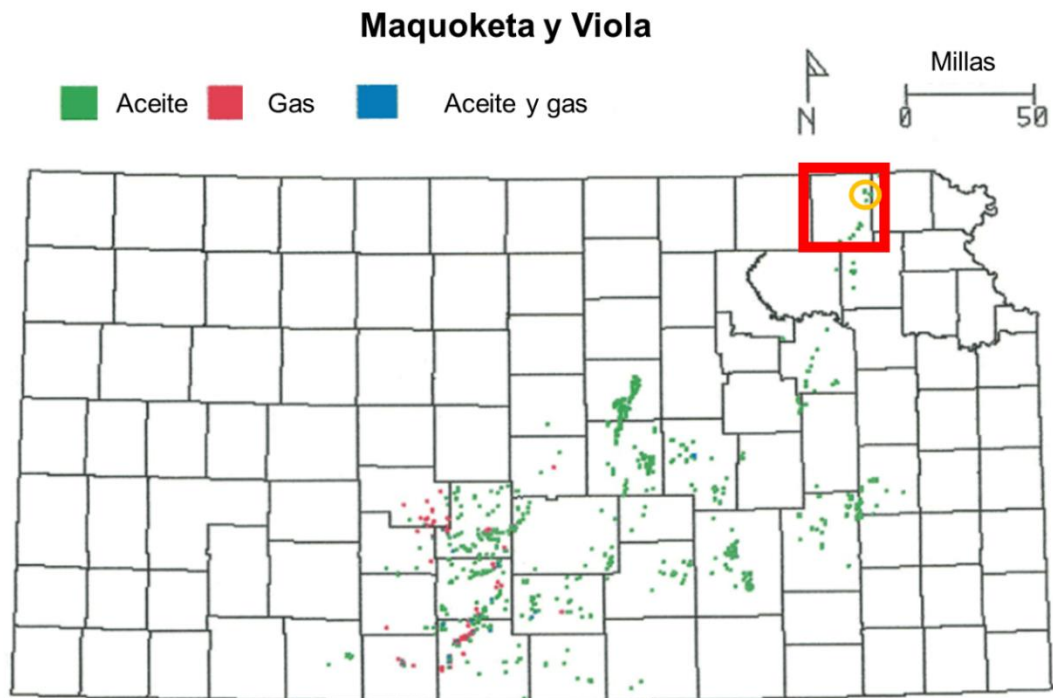


Figura 4.12. Zonas Productoras Pertenecientes a los Grupos Geológicos Maquoketa y Viola. Modificado de KGS.

La caliza del grupo geológico Viola puede encontrarse a lo largo de todo el estado a excepción del noroeste, se compone de calizas de grano grueso a fino y dolomitas que contienen una cantidad variable de esquisto (Bornemann, 1983). La caliza dolomitizada caracteriza la zona sur-central de Kansas, sin embargo, al noreste hacia Forest City y al este de la cuenca Salina en su mayoría se encuentra dolomita. Los tipos de porosidad varían, pero la intergranular, vugular, móldica y por fractura, pueden ocurrir (Caldwell & Boeken , 1985), esta columna estratigráfica se presenta en la **Figura 4.13**.









Tiempo Unidad Estratigráfica		Roca-Unidad Estratigráfica		
Sistema	Serie	Basado en correlación a las secciones superficiales.		Basado en el uso común por los geólogos petroleros de Kansas.
Ordovícico	Superior	Lutita Maquoketa		Lutita Maquoketa
		Caliza Viola		Caliza Viola
	Medio	Grupo Simpson		Grupo Simpson
	Inferior	Grupo Arbuckle		Grupo Arbuckle
Cámbrico	Superior	Dolomita Bonneterre		Grupo Arbuckle
		Arenisca Lamotte		
Precámbrico		Precámbrico		Precámbrico

Figura 4.13. Columna Estratigráfica del Grupo Viola y Maquoketa. Modificado de KGS.

En el embalse Hugoton la formación Viola se encuentra por encima de las rocas del mississippiico y de algunas otras rocas más antiguas hacia el este del embalse (Merrian & Atkinson, 1955) y también hacia el oeste (Maher & Collins, 1949). En el Suroeste de Kansas donde es difícil discernir de las rocas Arbuckle, la formación Viola va de grosores de 0 hasta 20 pies en los flancos del levantamiento de Kansas Central hasta más de 200 pies en las zonas más profundas del embalce de Hugoton cerca de la línea estatal con Colorado.

Los grupos Viola y Maquoketa no son los mayores productores de aceite del área Midcontinent, pero la producción más significativa de estas unidades ocurre en Kansas. Los campos más grandes en la cuenca de Forest City son en su mayoría trampas estructurales que producen de Viola. Estos campos incluyen a McClain, McClain suroeste, entre otros.

- Para el octavo pozo no se cuenta con la totalidad de la información de los estudios de núcleos para las distintas profundidades muestreadas, pero los atributos restantes están completos. Este es el sujeto que se separa para validar las herramientas. En la **Figura 4.14** se muestra su ubicación en el límite este del campo Hugoton Gas en el condado de Haskell, las muestras y mediciones de éste se tomaron de la formación Morrow.

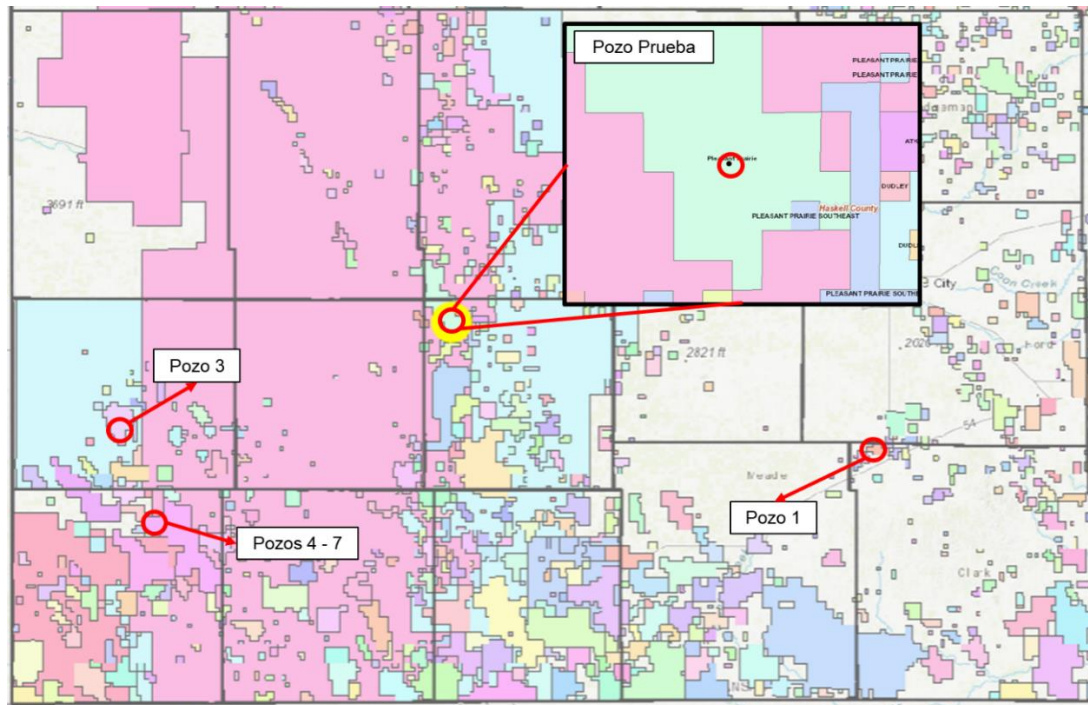


Figura 4.14. Ubicación del Pozo de Prueba.

4.3 Primer acercamiento a la Matriz de Datos

Con la finalidad de asegurar que la base de datos con la que se entrena a la red neuronal sea la mejor posible, la matriz de datos se somete a un preprocesamiento y precalificación mediante la aplicación de estadística básica y ciencia de datos.

El propósito de realizar este proceso es determinar si todas las instancias de la base son útiles para explicar el fenómeno que se está estudiando, así como conocer si el comportamiento de los atributos trabajados es correcto y se encuentra dentro de los límites teóricos y físicos posibles. En este trabajo se hará uso de las siguientes herramientas de estadística básica:

- Desviación estándar.
- Histogramas.
- Rango intercuartil de la variable.
- Valores máximos y mínimos
- Análisis de dispersión de los datos con gráficos de cajas y bigotes.

La matriz consta de nueve atributos, uno de los cuales es la propiedad que se busca aproximar (permeabilidad):

1. Profundidad [ft].
2. ILD [ohm-m].
3. GR [GAPI].
4. NPHI [fracción].
5. Porosidad [%].
6. Saturación de aceite [% del volumen poroso].
7. Saturación de agua [% del volumen poroso].
8. Densidad de grano [g/cc].
9. Permeabilidad [md] → Salida.

Se cuenta con 197 instancias que son revisadas para determinar que no existan problemas con los datos atípicos o comportamientos fuera de lo *natural*. Este primer proceso se realizó mediante la revisión de histogramas y el uso de gráficos de cajas y bigotes, dichos diagramas se muestran en las **Figuras 4.15** a la **Figura 4.22**.

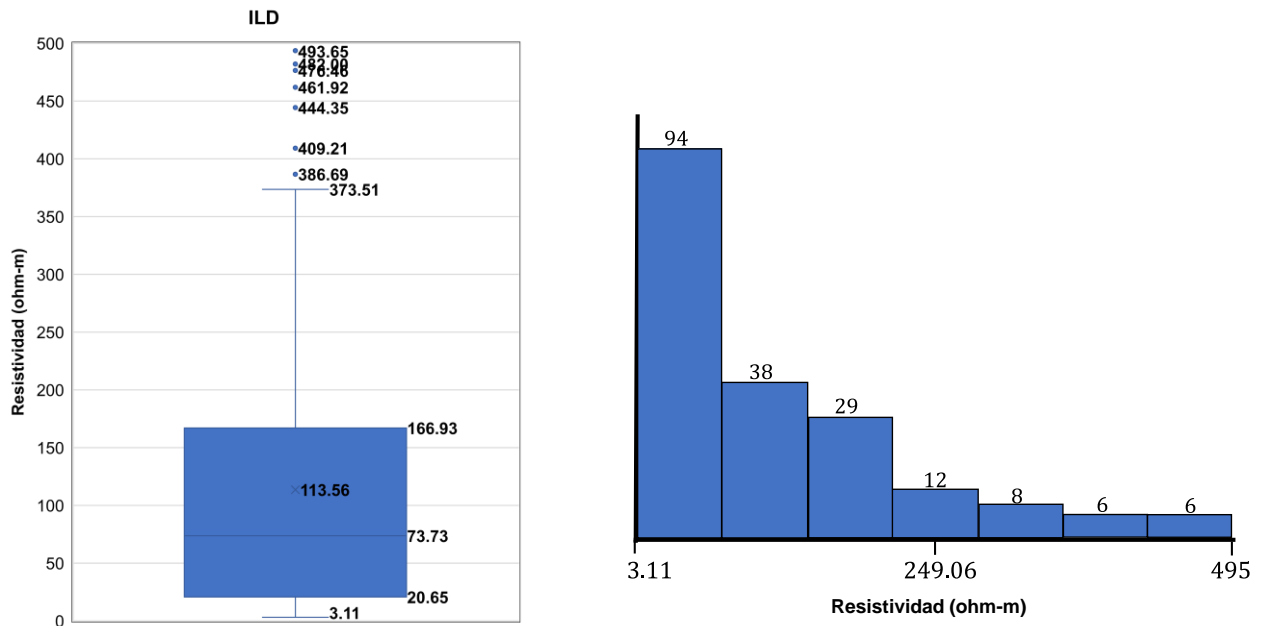


Figura 4.15. Boxplot (izquierda) e Histograma (derecha) del Registro ILD.

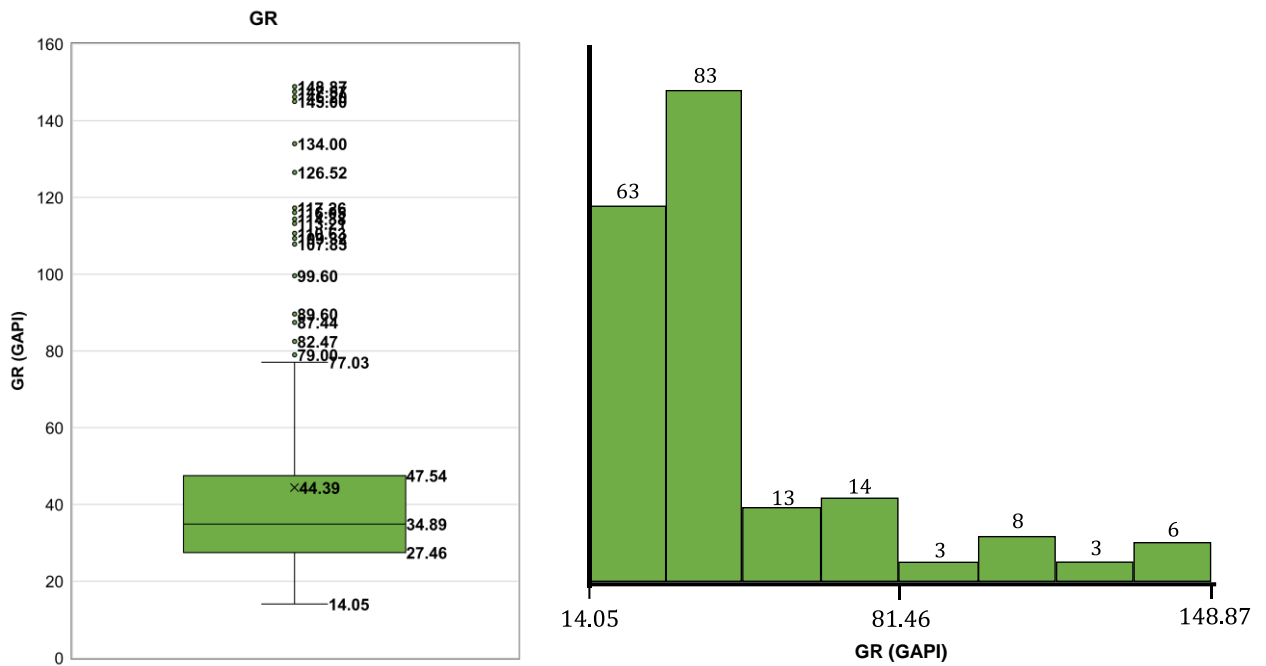


Figura 4.16. Boxplot (izquierda) e Histograma (derecha) del Registro GR.

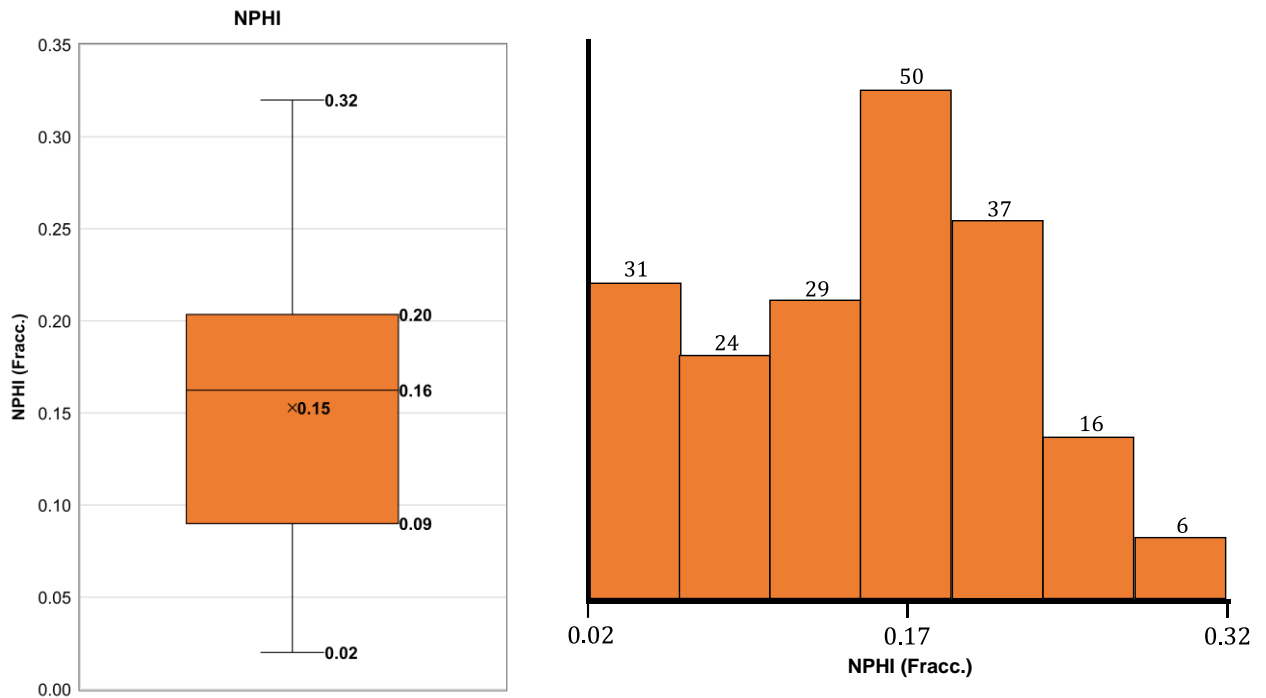


Figura 4.17. Boxplot (izquierda) e Histograma (derecha) del Registro NPHI.

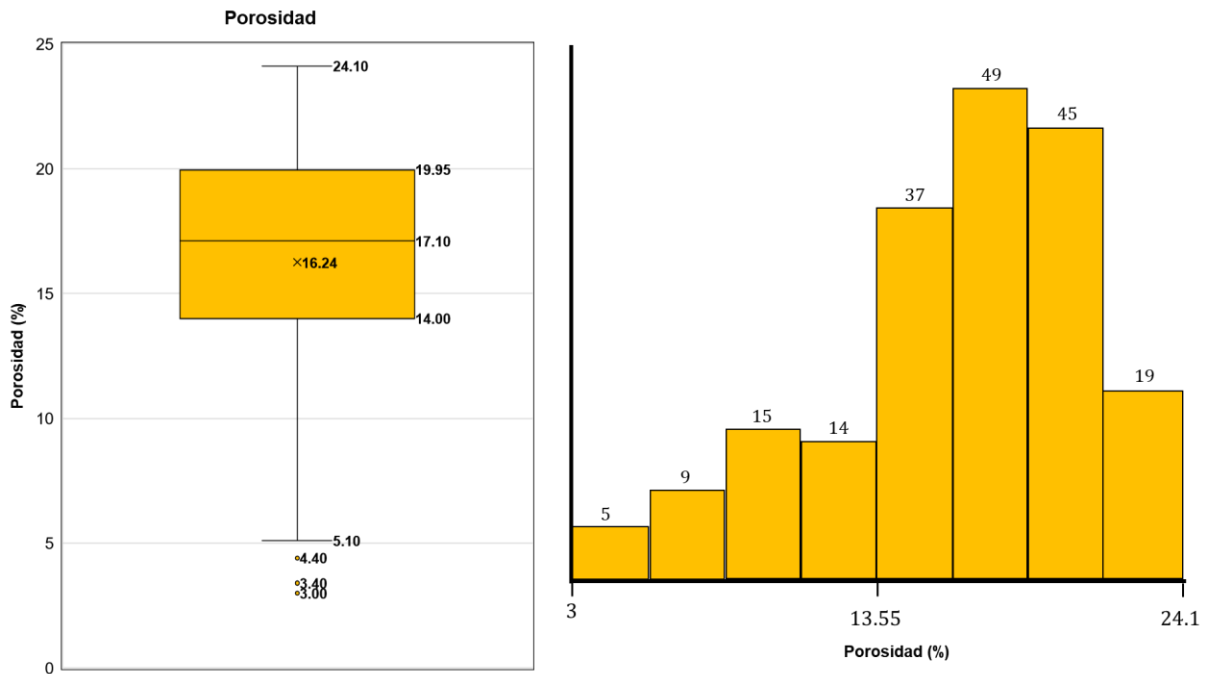


Figura 4.18. Boxplot (izquierda) e Histograma (derecha) de la Porosidad.

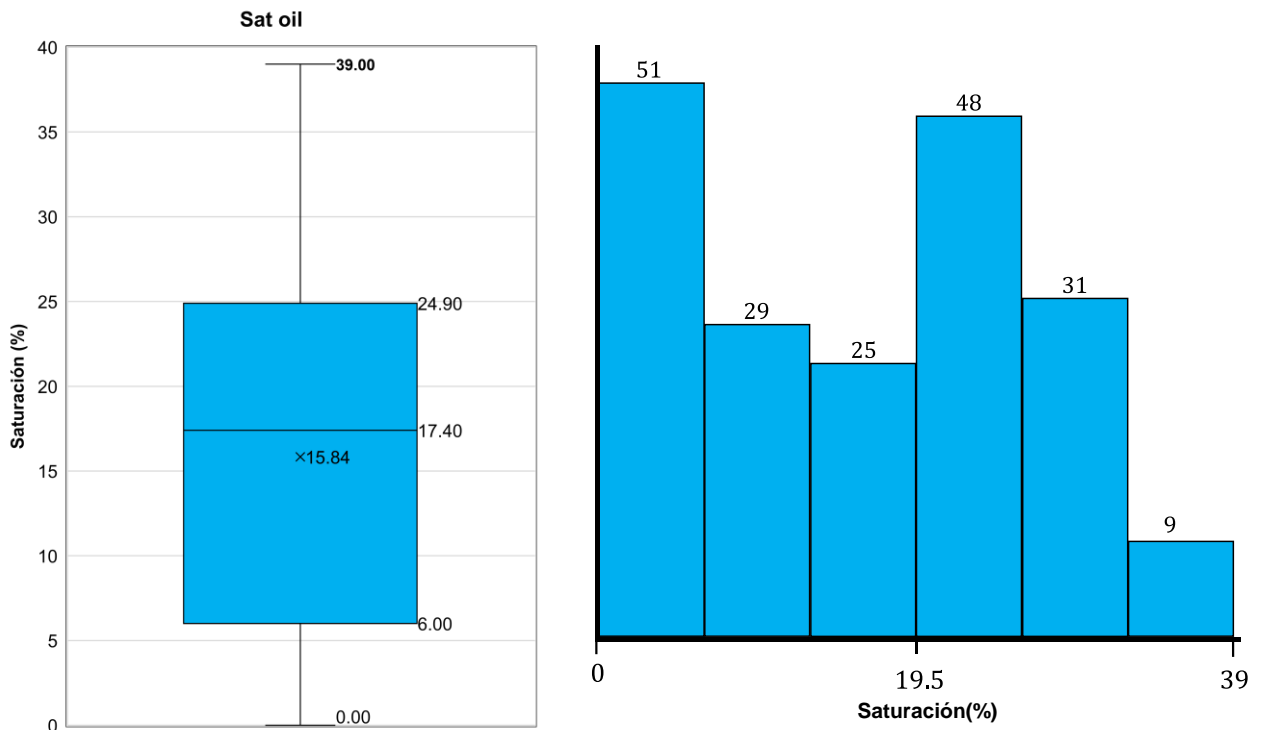


Figura 4.19. Boxplot (izquierda) e Histograma (derecha) de la Saturación de Aceite.

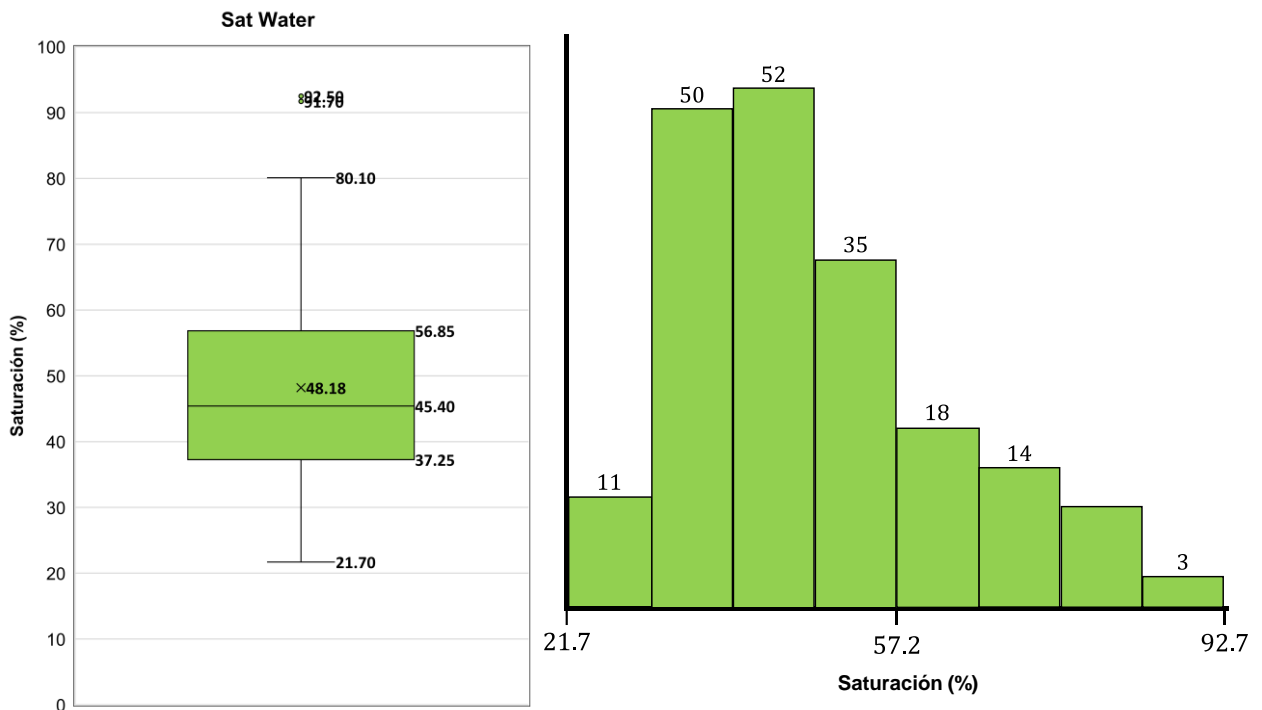


Figura 4.20. Boxplot (izquierda) e Histograma (derecha) de la Saturación de Agua.

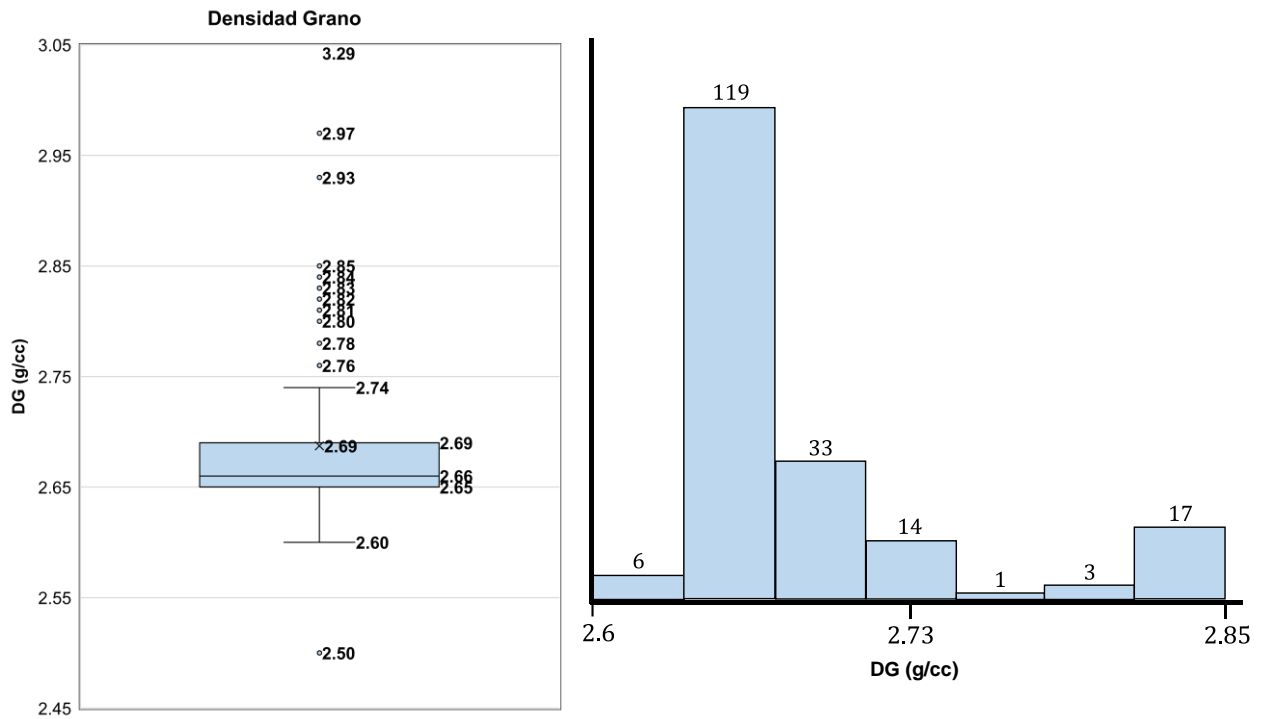


Figura 4.21. Boxplot (izquierda) e Histograma (derecha) de la Densidad de Grano.

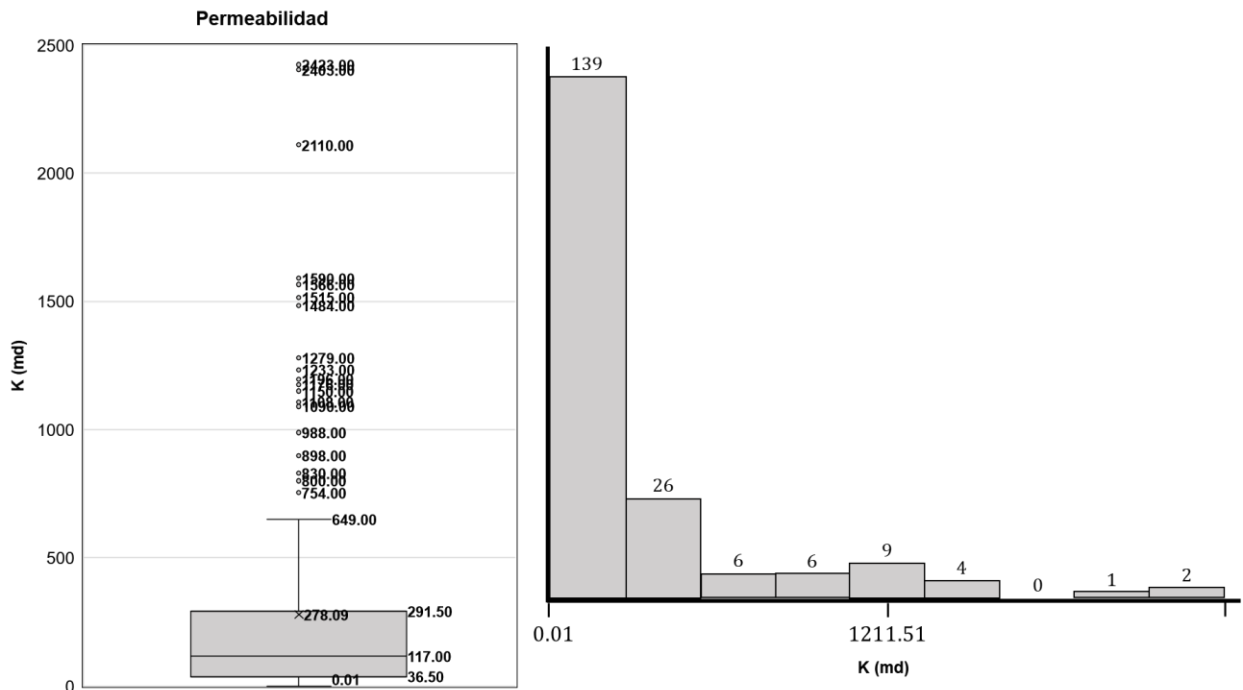


Figura 4.22. Boxplot (izquierda) e Histograma (derecha) de la Permeabilidad.

Gracias a este primer acercamiento a la matriz de datos y a la aplicación de estadística básica se decide sobre su posible reducción por comportamientos anormales. Se revisa entonces la información litológica y estratigráfica de las zonas donde se encuentran los pozos y de donde fue extraída la información.

Valores del Registro de Inducción (ILD)

La resistividad es la habilidad de un material para impedir el flujo de corriente eléctrica a través de él. La unidad es ohm-m. La matriz de roca, el aceite y el gas son aislantes eléctricos. Éstos no conducen el flujo de corriente eléctrica, además, se dice que sus resistividades son infinitas. Por su lado el agua conducirá la electricidad dependiendo de su salinidad, lo que implica que cualquier flujo de corriente a través de una formación toma lugar en el agua de formación y no en los hidrocarburos o la matriz de roca (Arroyo Carrasco, 2007).

Ya que el aceite y el gas no conducen la corriente eléctrica, es imposible distinguirlos de la matriz de roca solo con base en la resistividad, pero debido a los antecedentes con los que se cuenta para este trabajo, se sabe que las profundidades estudiadas y muestreadas corresponden a la formación productora para cada pozo, por lo que las altas resistividades aquí observadas serían efecto de la saturación de aceite en el medio.

Como regla general, las resistividades en formaciones arenosas caen en el rango de 0.2 y hasta 1000 ohm-m, mientras que para las formaciones calcáreas las resistividades pueden ser del orden de 100 a 40,000 ohm-m. Cabe aclarar que la herramienta resistiva utilizada tiene la capacidad de medir un máximo de 500 ohm-m (Arroyo Carrasco, 2007).

Valores del Registro de Rayos Gamma (GR)

La respuesta de una herramienta de este tipo depende del contenido de arcilla de una formación, esto se debe a que los elementos radiactivos tienden a concentrarse en las arcillas. Las formaciones limpias usualmente tienen un bajo nivel de contaminantes radiactivos. Según los datos con los que se cuenta el registro tiene una escala que va de 0 a 150 GAPI, donde las areniscas pueden marcar valores que van desde los 15 GAPI hasta alrededor de los 50 GAPI según estas sean limpias o tengan una mayor cantidad de lutitas pudiendo llegar a valores de 90 GAPI o más aquellas rocas con completamente

formadas por lutitas según su nivel de materia orgánica. Es importante resaltar que, en las rocas carbonatadas, podría observarse esta misma tendencia (KGS, 2020).

Valores del Registro Neutrón Compensado (NPHI)

Esta herramienta utiliza una fuente radiactiva emisora de neutrones rápidos y dos detectores. La medición se basa en la relación de conteos de estos detectores, reflejando la forma en la cual la densidad de neutrones decrece con respecto a la distancia de la fuente y esto depende del fluido (índice de hidrógeno) contenido en los espacios porosos de la roca. Por lo tanto, el flujo de neutrones pierde energía al pasar por la formación, cuando esta alta energía con la que son emitidos disminuye por colisiones con elementos que tienen una masa atómica muy parecida a la del neutrón, en este caso al colisionar con el hidrógeno presente en las moléculas de agua y de hidrocarburos. La herramienta resulta más efectiva en formaciones no arcillosas con líquido en los poros cuyos rangos de porosidad sean menores a 25% (Arroyo Carrasco, 2007). El rango para el registro usado va de 0 a 45% medido en fracción.

Valores de Porosidad

Esta medición fue obtenida de análisis en laboratorios y basándose en los antecedentes geológicos mostrados con anterioridad las formaciones productoras están clasificadas como rocas areniscas o calizas dolomitizadas con granos esqueléticos disueltos, lo que nos muestra que existen rangos de porosidad que van de 25 al 40% en el caso de las areniscas y del 0 al 70% en el caso de las calizas (Alonso, 2006).

Valores de Densidad de Grano (DG)

En el caso de la densidad de grano, es un parámetro igualmente obtenido de laboratorio y para las areniscas existe un rango de valores entre 2 y 2.8, mientras que para la caliza dolomitizada pueden observarse valores de densidad de grano entre 2.4 y 2.9 (**Figura 4.23**).

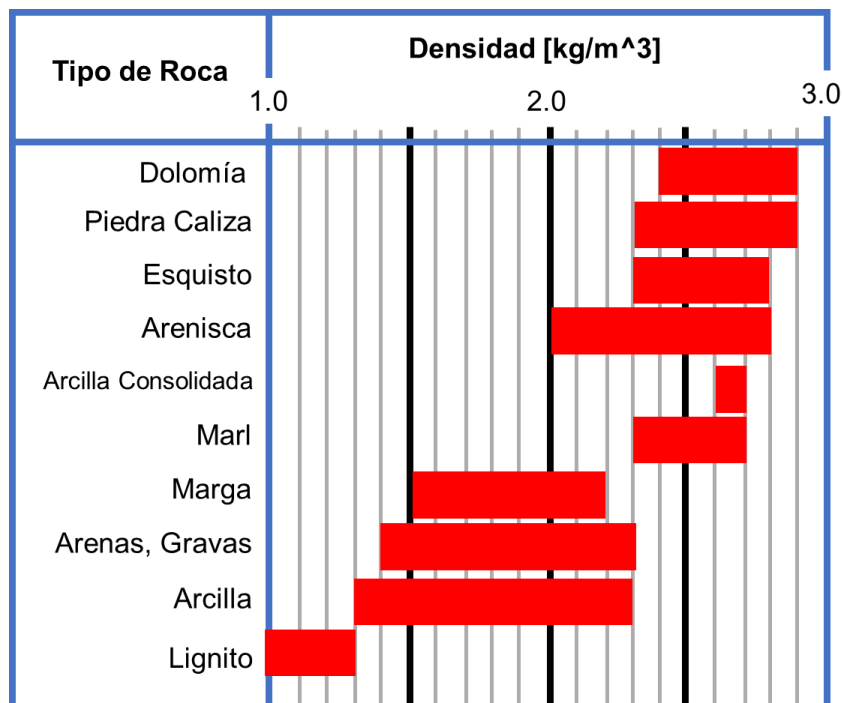


Figura 4.23. Densidad de las Principales Rocas Sedimentarias. Modificada de (Wohlenberg, 1982).

Valores de Permeabilidad (K)

Para el caso de la permeabilidad, la **Tabla 4.1** resume los rangos esperados de las formaciones rocosas encontradas según el estudio geológico del área donde se encuentran los pozos.

Tabla 4.1. Rangos de permeabilidad de rocas. Modificada de (Fetter, 2001).

Denominación	Permeable				Semi-permeable			Impermeable			
Roca	Rocas muy fracturadas				Roca Petrolífera			Piedra Arenisca	Dolomita		
K (md)	10 ⁸	10 ⁷	10 ⁶	10 ⁵	10 ⁴	1000	100	10	1	0.1	0.01

De los valores anteriormente mostrados es importante recordar que la permeabilidad suele aumentar por la existencia de fallas, grietas, porosidad de naturaleza secundaria y otros efectos estructurales como es el caso de las formaciones carbonatadas

pertencientes al área del campo Hugoton Gas y de los grupos geológicos Viola y Maquoketa.

Tomando en consideración lo anterior, cuatro instancias fueron eliminadas de la matriz de datos original, dejando un total de 193 (de 197 lo que da una idea de la consistencia del repositorio y la confianza sobre los valores).

La **Tabla 4.2**, presenta valores máximos, mínimos, media, desviación estándar, rango intercuartil y tipo de variable para cada uno de los atributos de la base de datos después de este primer proceso, pues es de gran importancia tener estos parámetros presentes para la siguiente etapa de análisis de los datos.

Tabla 4.2. Valores de Estadística Básica de la Base de Datos.

Atributo	Máx.	Min.	Med.	DE.	Rango IC	Tipo
Registro ILD	495	3.11	115.052	118.41	352.86	Numérica
Registro GR	148.87	14.05	44.39	28.76	20.1	Numérica
Registro NPHI	0.32	0.02	0.153	0.068	0.11	Numérica
Porosidad	24.1	3	16.39	4.58	13.89	Numérica
Sat. Aceite	39	0	15.93	11.17	18.9	Numérica
Sat. Agua	92.7	21.7	48.271	14.05	19.6	Numérica
Densidad Grano	2.85	2.6	2.68	0.056	0.04	Numérica
Permeabilidad	2423	0.01	283.55	430.81	255	Numérica

Finalmente, se presentan los comportamientos de los atributos (entradas - salida) para los 7 pozos en las **Figuras 4.24 a 4.30**.

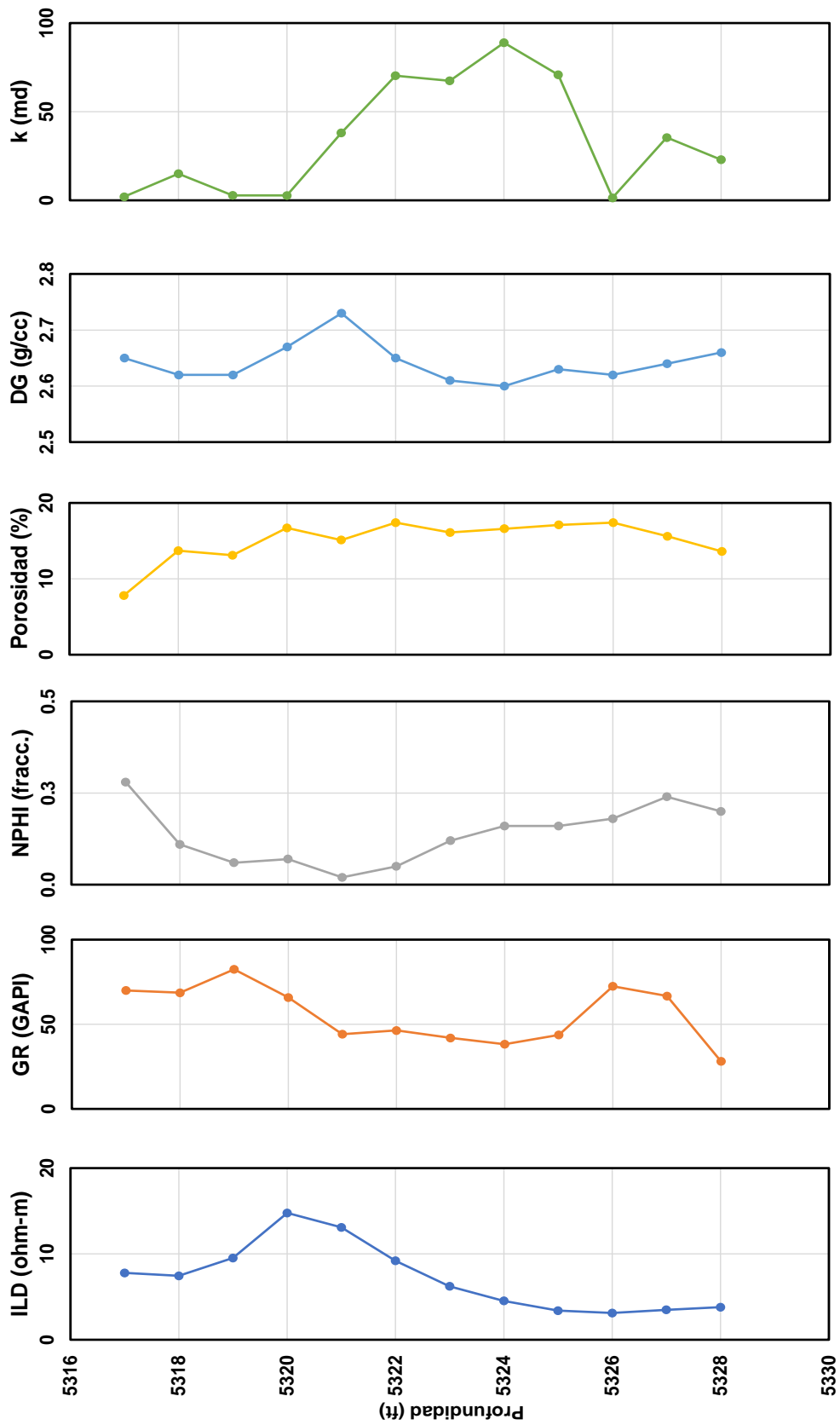


Figura 4.24. Comportamiento de los Atributos del Pozo 1.

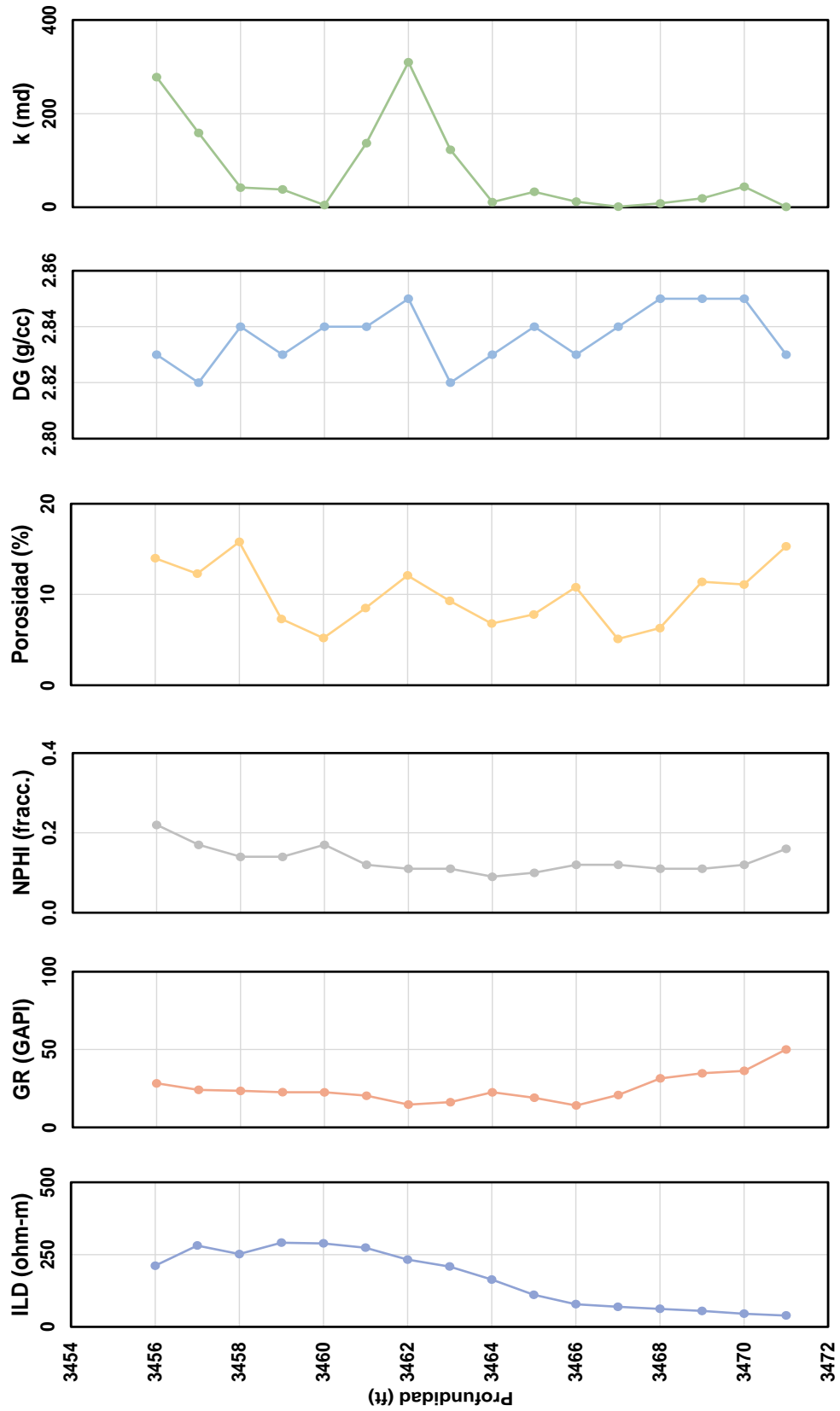


Figura 4.25. Comportamiento de los Atributos del Pozo 2.

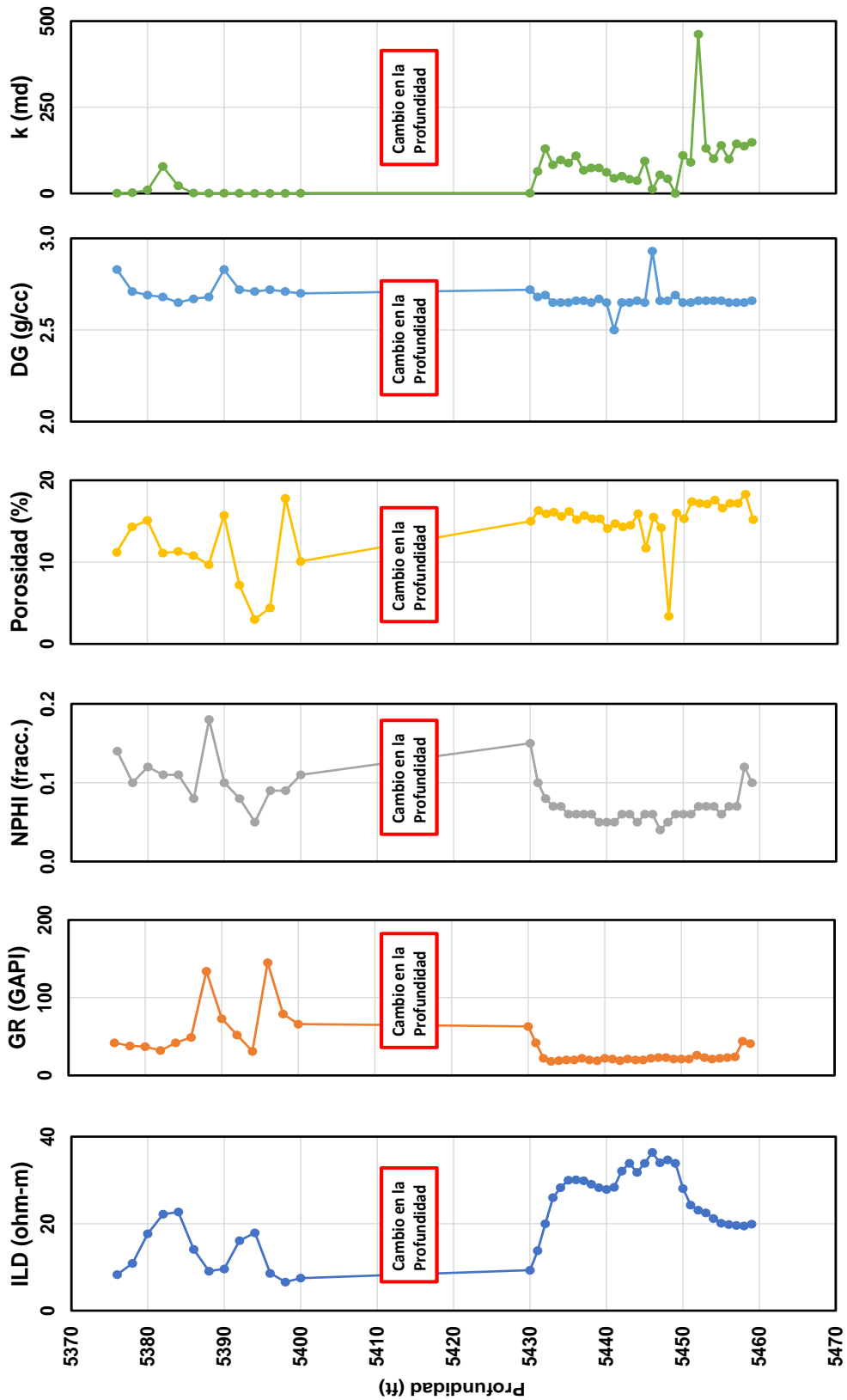


Figura 4.26. Comportamiento de los Atributos del Pozo 3.

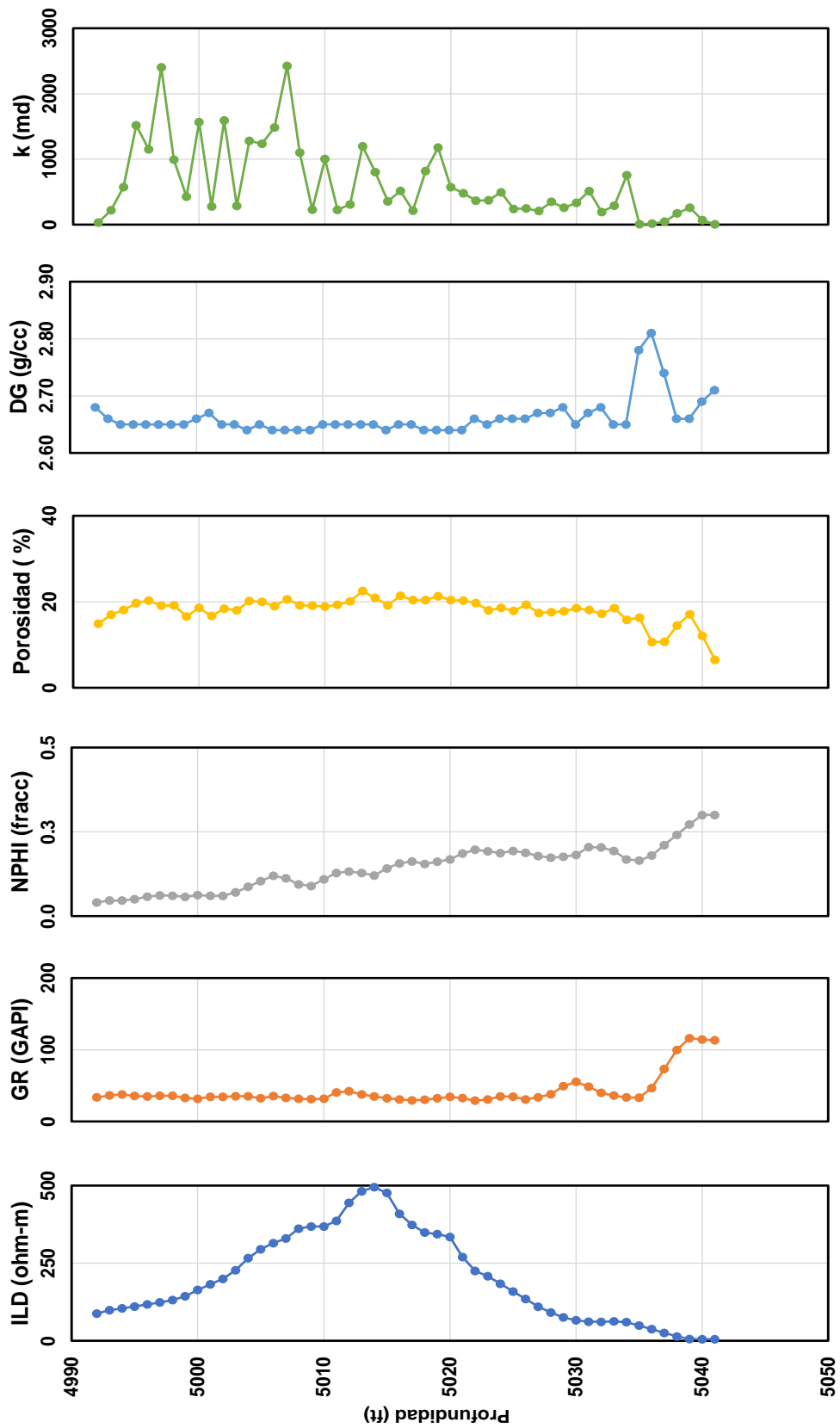


Figura 4 27. Comportamiento de los Atributos del Pozo 4.

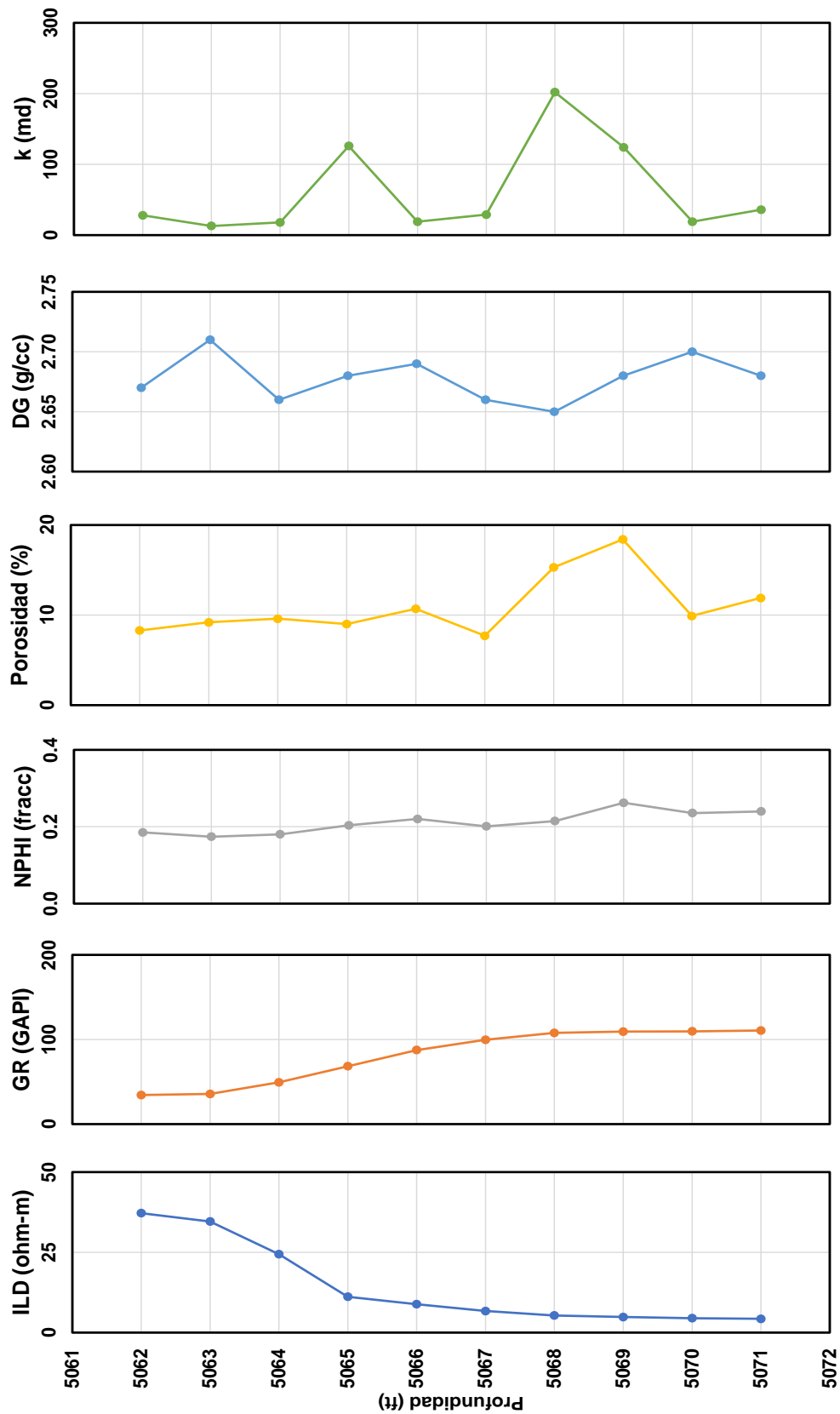


Figura 4.28. Comportamiento de los Atributos del Pozo 5.

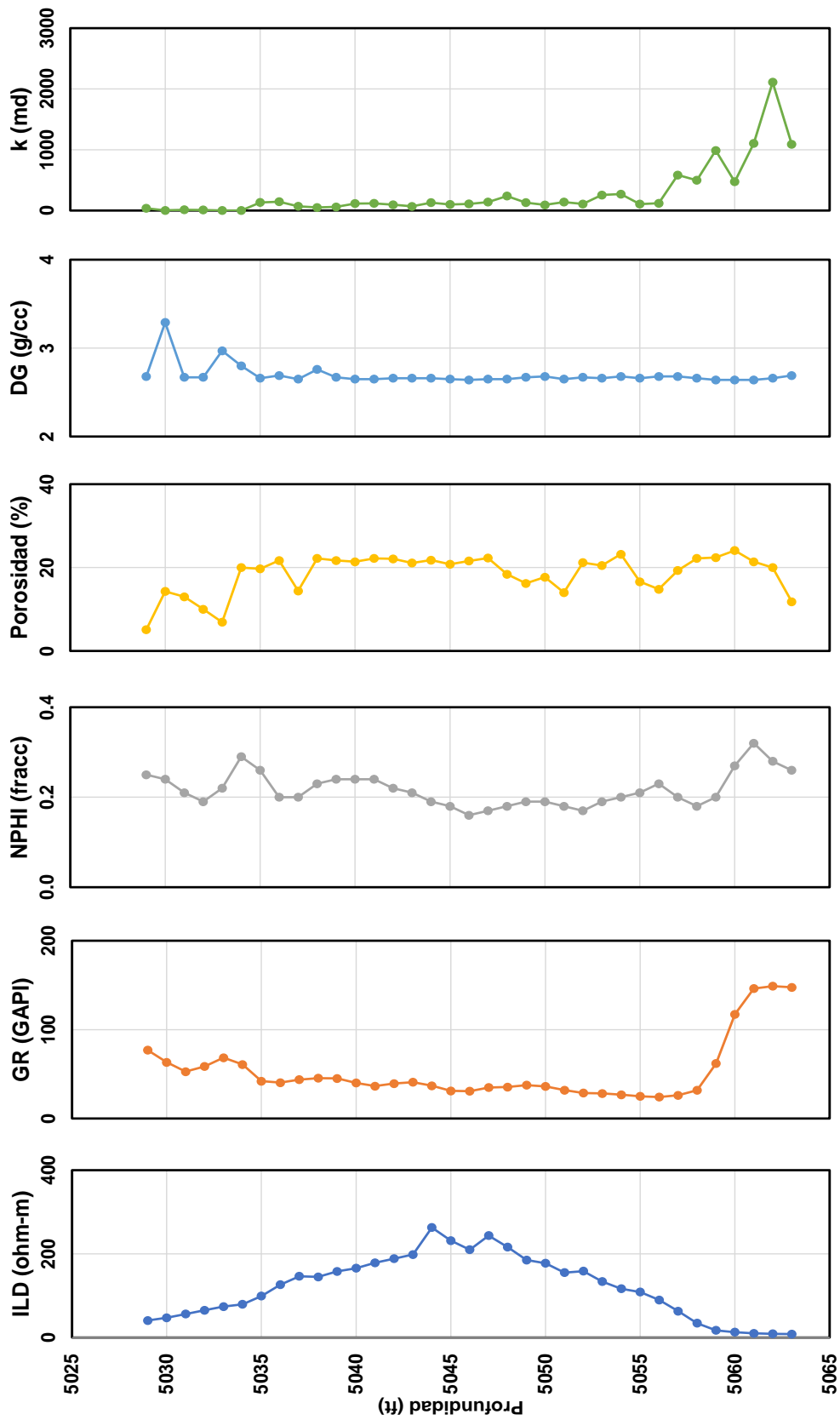


Figura 4.29. Comportamiento de los Atributos del Pozo 6.

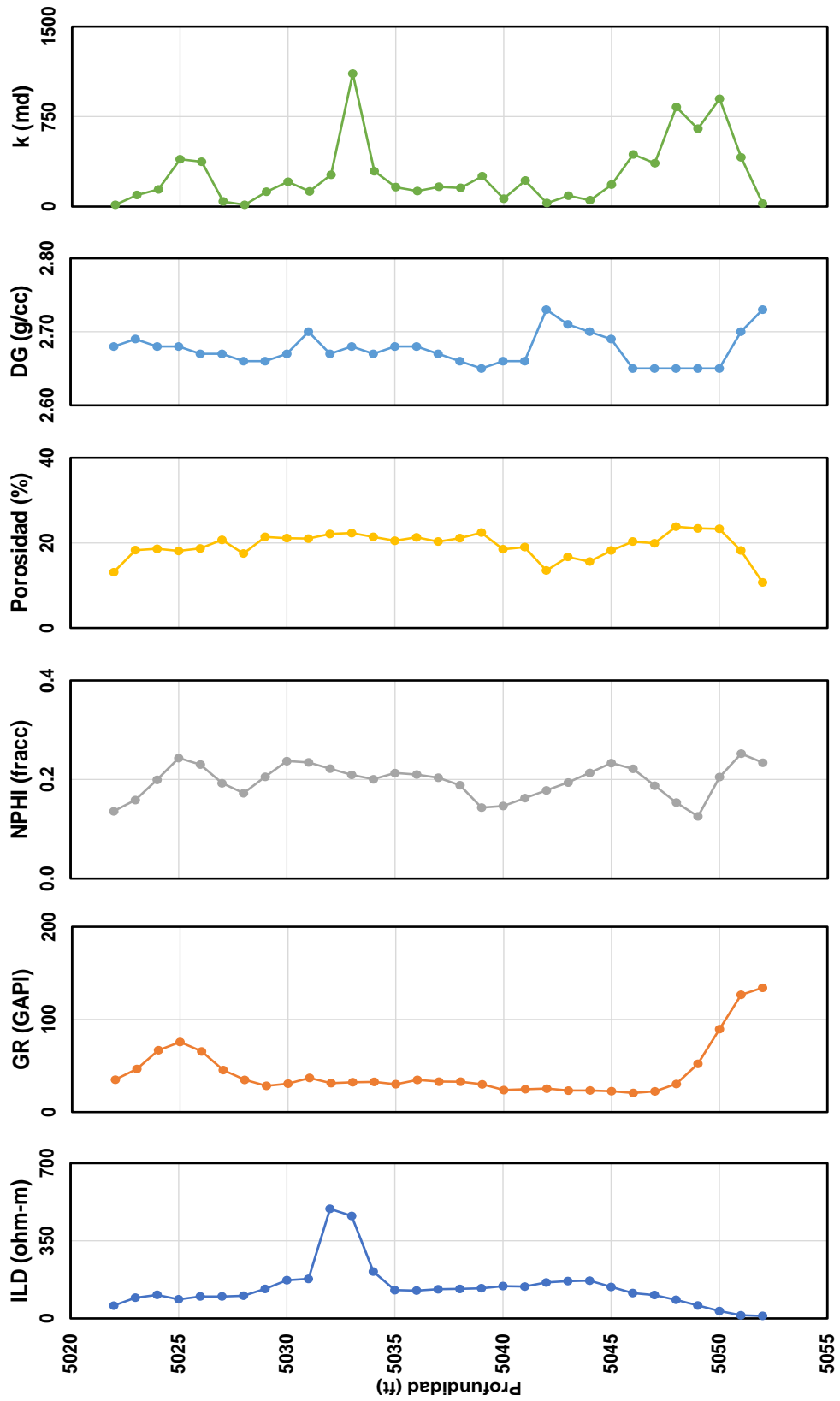


Figura 4.30. Comportamiento de los Atributos del Pozo 7.

4.4 Ciencia de Datos sobre la Matriz Reducida

En cuanto a las herramientas de Ciencia de Datos, útiles para la precalificación de la base, se optó por el uso de las siguientes:

- Prueba gamma.
- Árbol de regresión (CART).
- Validación cruzada (k-folds cross-validation), usada dentro del análisis del árbol de regresión.
- Análisis factorial.

4.4.1 Resultados: Árbol de Regresión

Se dispuso que ocho atributos cumplieran la función de ser usados como entradas del modelo y uno como salida (permeabilidad). Se utilizó el algoritmo M5P para árboles de regresión con una partición del 85% de las 193 instancias para ser usadas en el entrenamiento, mientras que el resto (30) fueron utilizadas en la prueba del modelo generado. El número de secciones del total de los datos que fueron escogidos para entrenamiento y prueba se optimizó mediante la aplicación de validación cruzada (k-folds cross validation).

En la **Figura 4.31**, se presenta el árbol de regresión generado. Se observa:

- El nodo director o raíz es el atributo porosidad, lo cual se espera por ser la variable que, entre las incluidas, tiene relación directa con la permeabilidad. La primera separación para porosidades menores o iguales a y mayores a 17.95 resulta ser una primera segmentación sobre regiones: entre los pozos pertenecientes al campo Hugoton Gas y aquellos de los campos de donde se produce la formación Morrow y Viola.

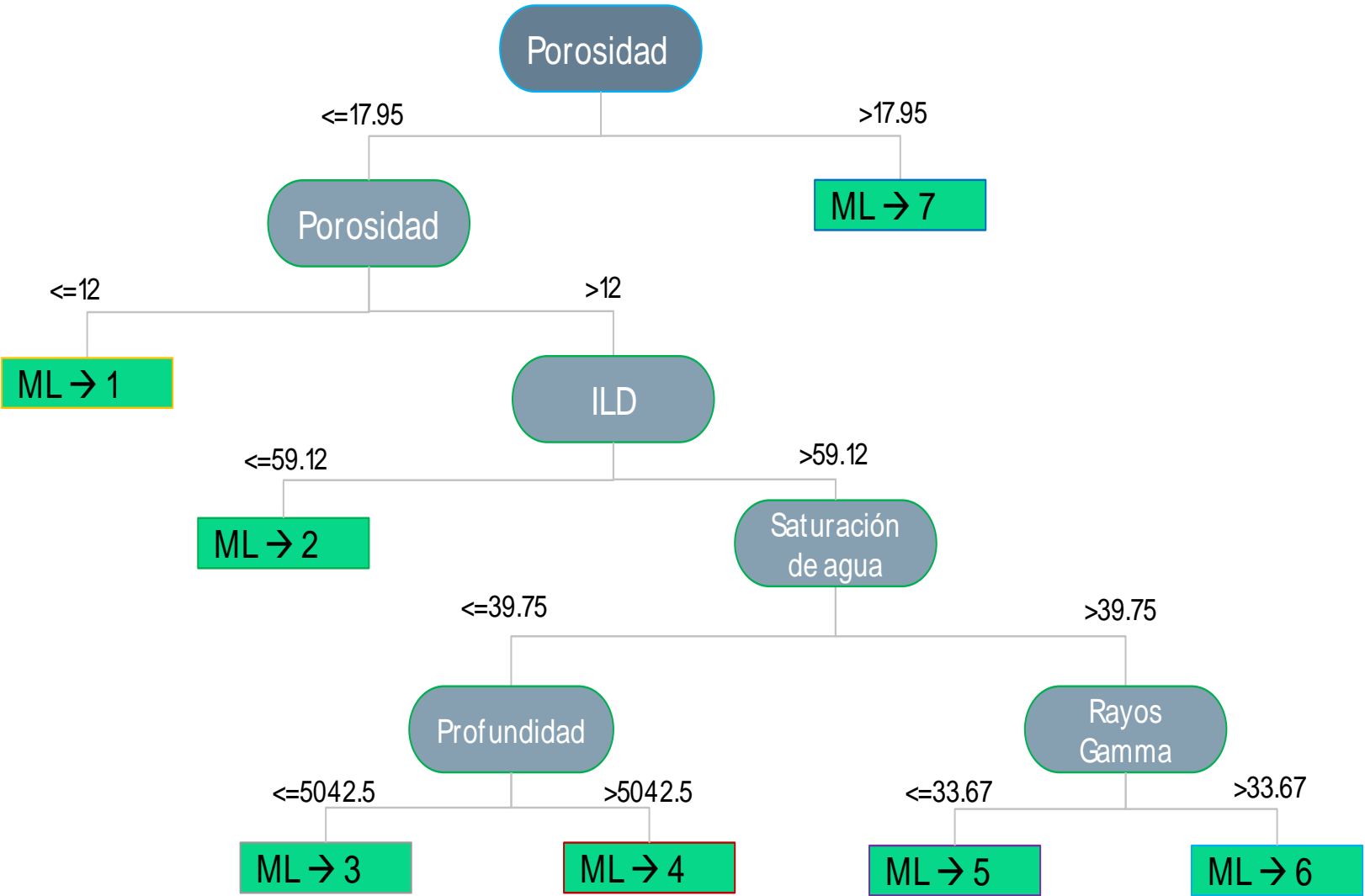


Figura 4.31. Resultado del Árbol de Regresión.

- Para los valores de porosidad mayores a 17.95 tres de los cuatro pozos en la sección S19-T31S-R40W pertenecientes a las formaciones del Morrow superior del campo Hugoton Gas, se asocian al nodo terminal con modelo lineal 7 (*ML7*), mientras que los pozos restantes de la misma zona junto con el pozo de la formación Viola forman parte de las muestras con porosidades menores a 17.95%.
- Un siguiente segmento en la porosidad se marca en 12%. Para instancias con porosidades menores o iguales a 12%, dos pozos, el encontrado en la formación Viola y el restante del conglomerado de cuatro encontrados en el condado de Morton (pozo 5), quedan separados. Es importante señalar que el pozo 5 está 30 pies más profundo que los demás pozos, ubicándose así en la parte inferior de la formación Morrow. En estos casos los valores de permeabilidad se estiman con el modelo lineal 1 (*ML1*).
- Para instancias con porosidad entre 12 y 17.95% la segmentación se produce con el registro ILD en la frontera 59.12 ohm-m, valor muy cercano a la diferencia de resistividades entre una matriz de arenisca y una de caliza, ambas conteniendo aceite. Casos con valores menores o iguales a 59.12 ohm-m están en el campo Arroyo, condado de Staton y el pozo del campo Norcan East ambos productores de la formación Morrow del Pensilvánico inferior y medio (sección de arenisca), tamaños de grano variado, así como porosidades y permeabilidades menores. Estimación de k con el modelo lineal 2 (*ML2*).
- Las muestras con resistividades mayores a 59.12, 20 instancias, son agrupadas a partir de la saturación de agua, profundidad y registro GR con modelos en hoja final *ML3, ML4*. La separación por valor del registro de rayos gamma asigna *ML5* y *ML6* en sus ramas izquierda y derecha, respectivamente.

A continuación, los 7 modelos lineales:

$$ML1 \rightarrow K = -0.047 * DEPTH + 0.2594 * ILD + 0.5208 * GR + 10.1934 * \phi - 1.2212 * S_o + 0.6101 * S_w - 430.2752 * DG + 1263.0942.$$

$$ML2 \rightarrow K = -0.0533 * DEPTH + 0.2884 * ILD + 0.5208 * GR + 29.5705 * \phi - 1.2212 * S_o - 0.7316 * S_w - 555.3257 * DG + 1419.2418.$$

$$ML3 \rightarrow K = 0.1799 * DEPTH - 0.3268 * ILD - 2.1528 * GR + 12.535 * \phi + 1.5581 * S_o + 8.2466 * S_w - 686.6432 * DG + 639.4716.$$

$$ML4 \rightarrow K = 0.339 * DEPTH - 0.3268 * ILD - 2.1528 * GR + 12.535 * \phi + 1.5581 * S_o + 8.2466 * S_w - 686.06432 * DG - 161.1335.$$

$$ML5 \rightarrow K = -0.0089 * DEPTH - .7665 * ILD - 3.9624 * GR + 12.535 * \phi + 1.0022 * S_o + 6.7193 * S_w - 686.6432 * DG + 1845.8941.$$

$$ML6 \rightarrow K = -0.0089 * DEPTH - 0.5531 * ILD - 3.8604 * GR + 12.535 * \phi + 1.0022 * S_o + 6.7193 * S_w - 686.6432 * DG + 1794.689.$$

$$ML7 \rightarrow K = -1.6944 * DEPTH - 0.5531 * ILD - 3.8604 * GR + 54.2952 * \phi - 24.1935 * S_o + 16.1558 * S_w - 333.9511 * DG + 8512.8226.$$

Las estimaciones de permeabilidad con cada modelo lineal dictan las adecuaciones al conjunto de datos para su posterior uso con redes neuronales. En la **Figura 4.32** se muestran las segmentaciones por instancia (punto evaluado en profundidad o núcleo) empezando por el nodo raíz, la porosidad. Las líneas verticales separan los valores de la primera bifurcación. Para valores mayores a 17.95% de porosidad se calcula el valor de k con *ML7*. En este caso una instancia del pozo 1 pertenece a este modelo lineal, resaltada con color azul y la etiqueta correspondiente.

Todos los valores menores o iguales a 17.95% pasan al siguiente nodo de decisión donde la porosidad una vez más vuelve a ser el atributo para evaluar con una bifurcación en 12% (esta se encuentra representada por la segunda línea punteada vertical, segunda de derecha a izquierda) en el mismo gráfico. Para valores menores o iguales a 12% de porosidad, la permeabilidad se calcula mediante el modelo lineal 1 (*ML1*). En este pozo una instancia usa esta regla, color naranja con su respectiva etiqueta.

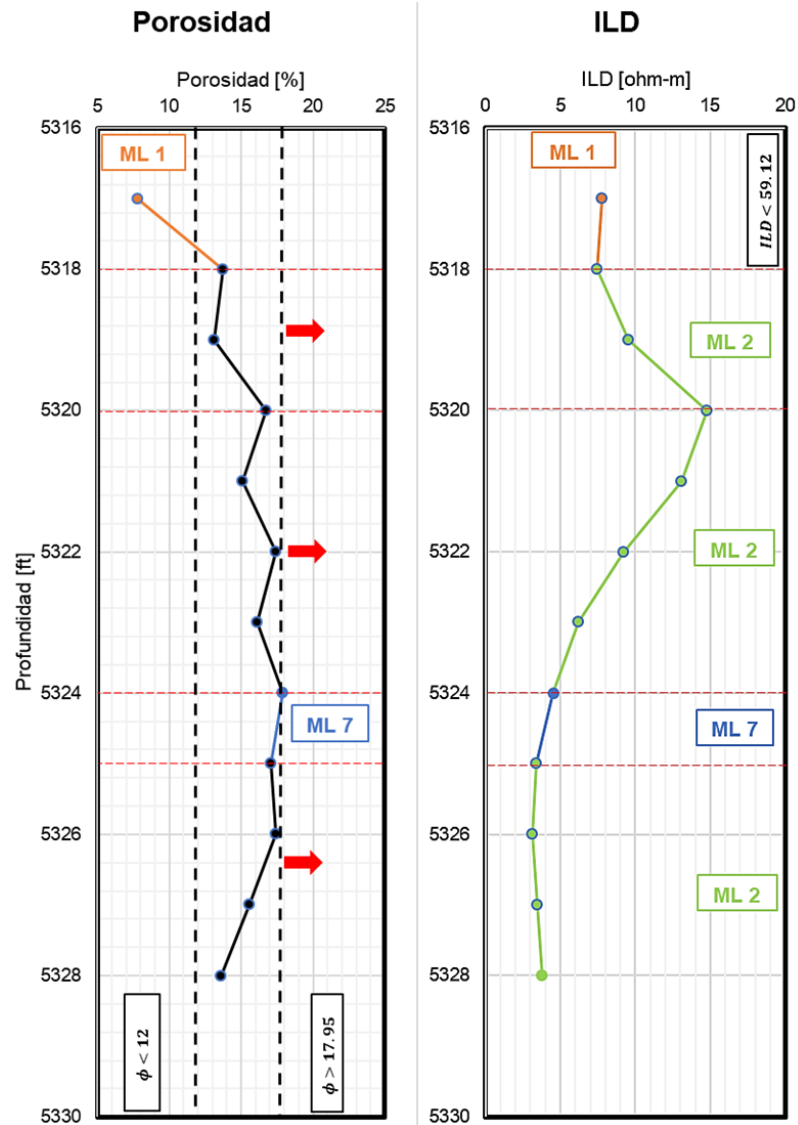


Figura 4.32. Procedimiento para el Pozo 1.

Ahora bien, para las instancias con valores de porosidad entre 12% y 17.95% se usa registro ILD (límite de 59.12 ohm-m) para los siguientes cálculos. En la **Figura 4.33** se comparan las permeabilidades estimadas con el árbol de regresión (línea gris) con las mediciones en estudio de núcleo (línea roja).

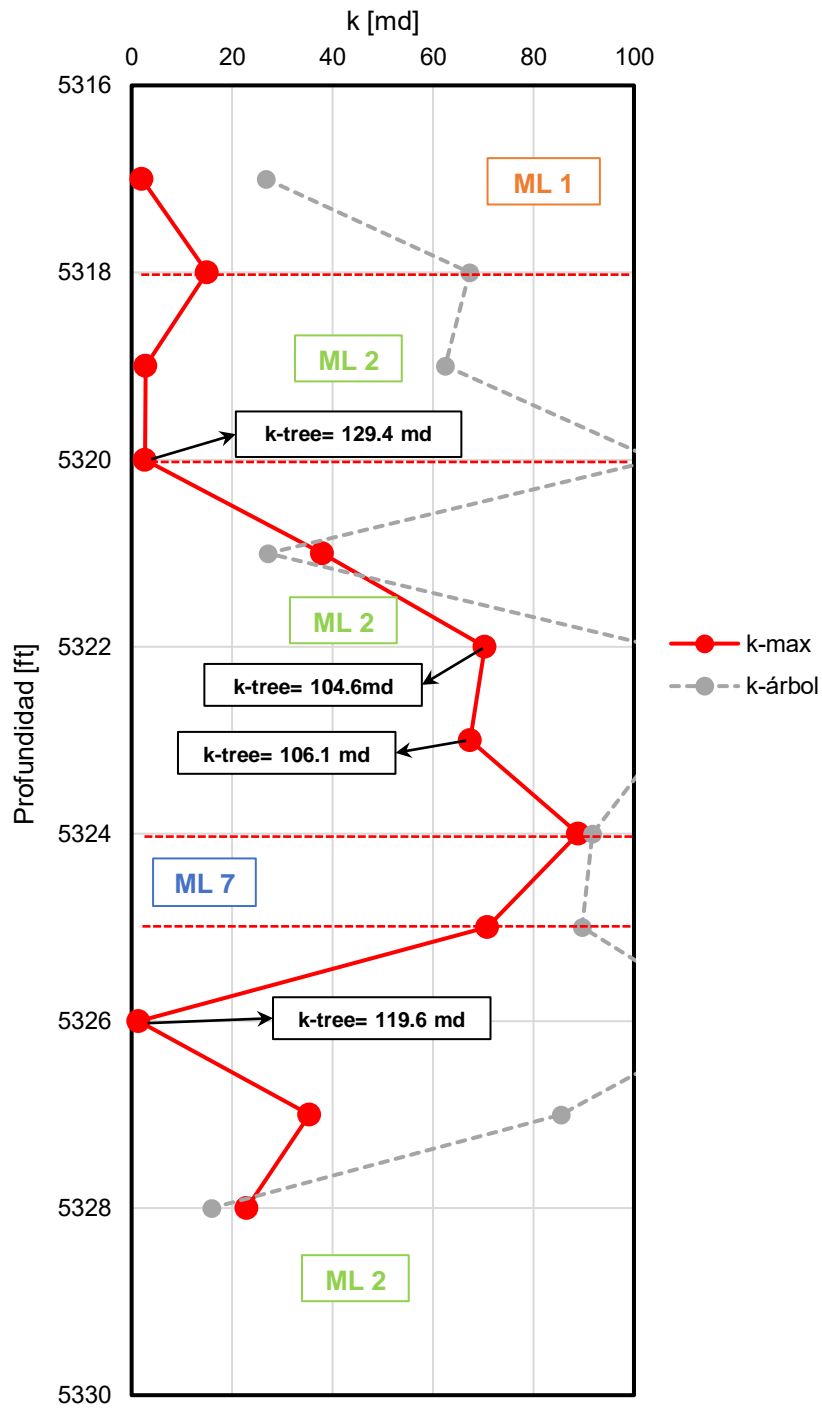


Figura 4.33. Permeabilidad Medida Contra Evaluada para el Pozo 1 (k -max: permeabilidad medida, k -árbol: permeabilidad estimada por el árbol).

El pozo 2 sigue un camino más largo en la estructura del árbol de regresión. En la **Figura 4.34** se muestra la segmentación por el procedimiento.

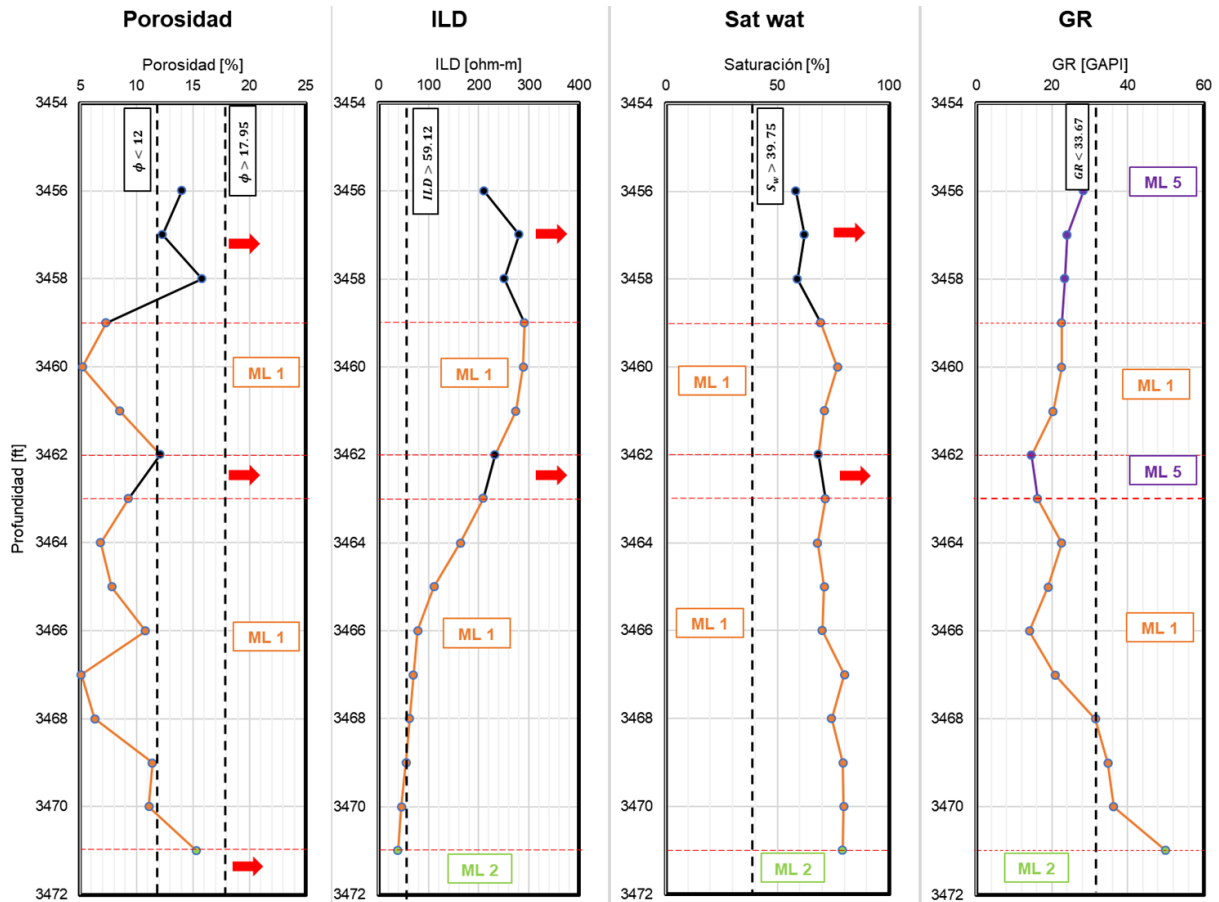


Figura 4.34. Procedimiento para el Pozo 2.

Para este pozo se evalúan aquellas instancias que tienen porosidades entre 17.95% y 12%, resaltadas con una línea negra en el gráfico de porosidad. Desde el segundo nodo de decisión (gráfico del registro ILD) una de las instancias se coloca en el nodo terminal del modelo lineal 2 (*ML2*). Para las otras instancias la saturación de agua (valor de partición en 39.75%) y registro de rayos gamma (valores frontera en GAPI) llevan al uso del modelo lineal 5 (*ML5*), correspondiente a la línea y etiqueta de color púrpura.

En la **Figura 4.35** se muestran los resultados de la aplicación del árbol a este pozo, donde la línea roja corresponde a las mediciones realizadas en un núcleo mientras que la línea gris son los valores de permeabilidad estimados.

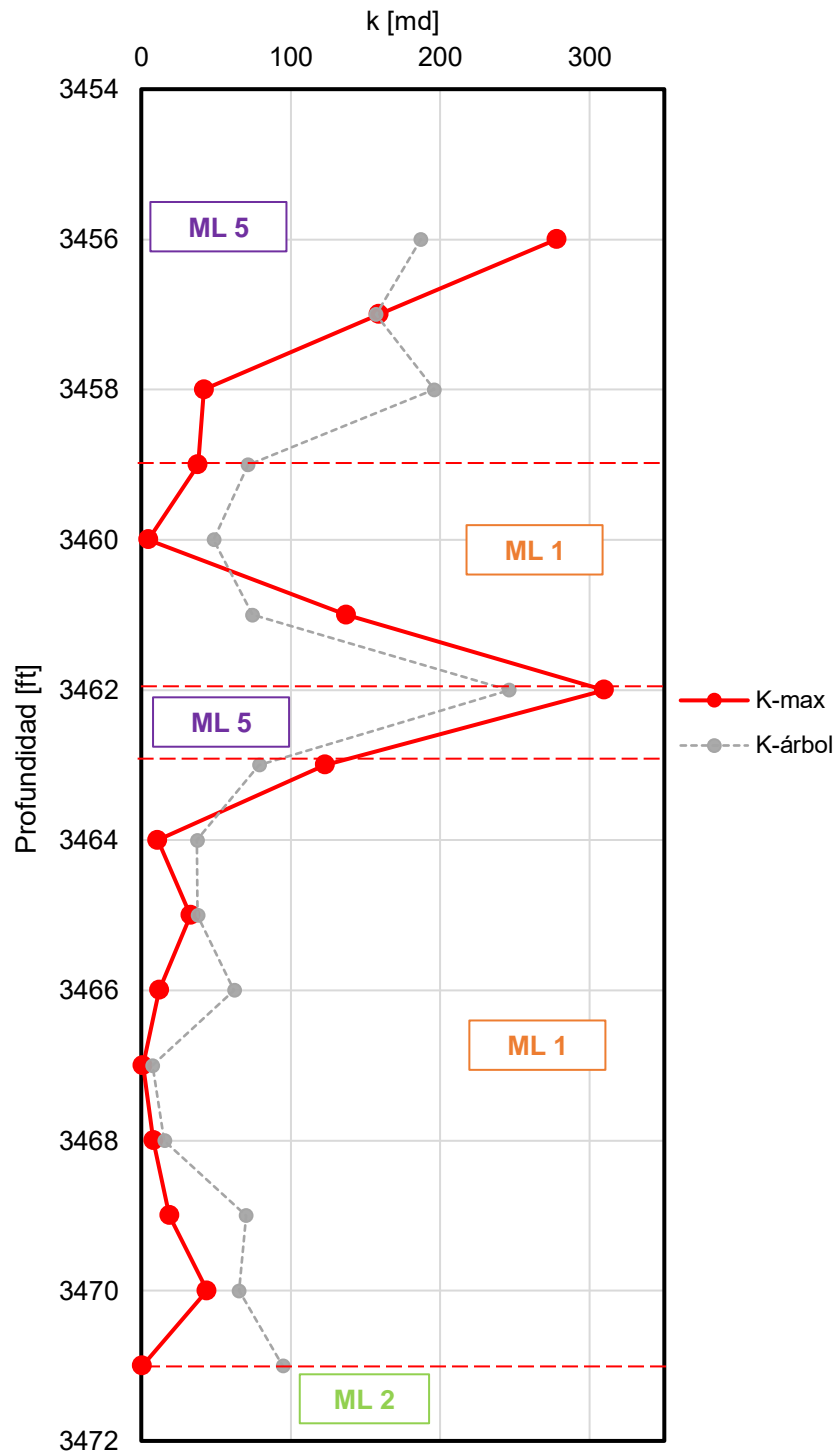


Figura 4.35. Permeabilidad Medida Contra Evaluada para el Pozo 2 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).

En la **Figura 4.36** se muestra el procedimiento aplicado al pozo 3.

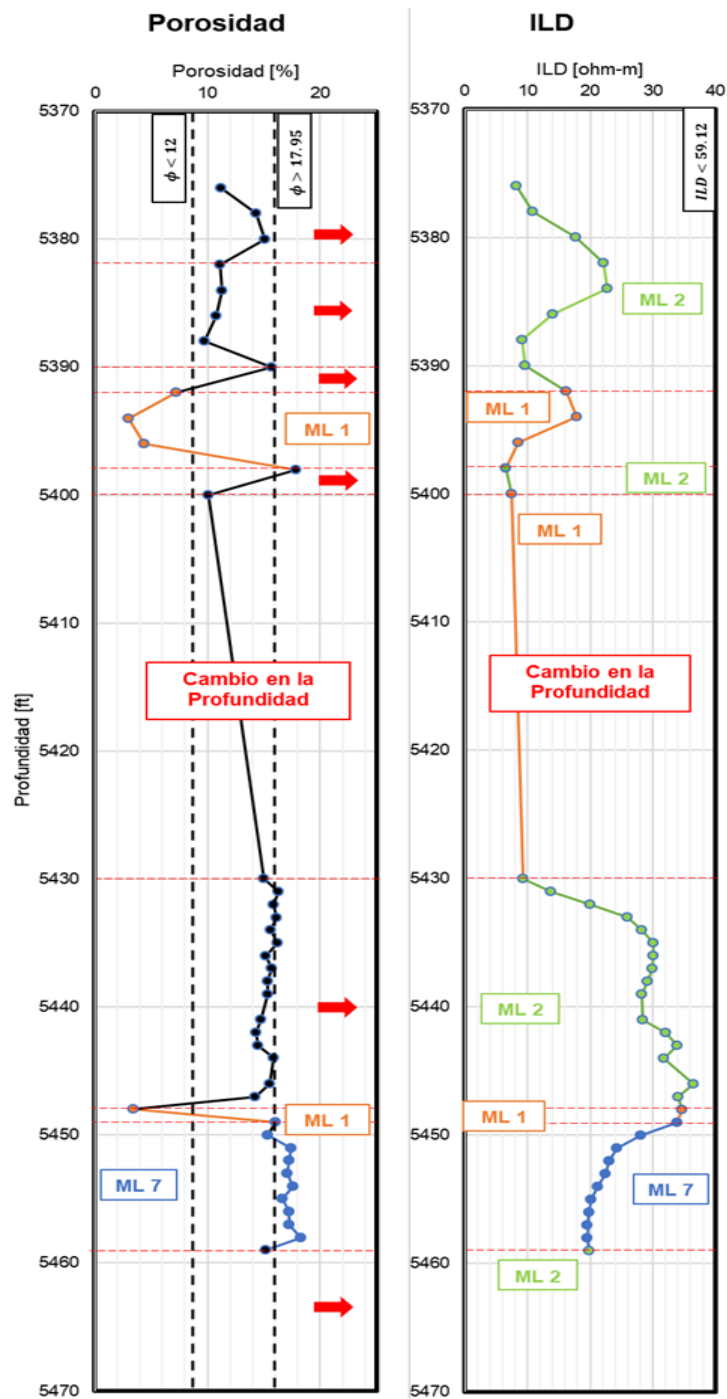


Figura 4.36. Procedimiento para el Pozo 3.

La **Figura 4.37** muestra las permeabilidades medidas contras las estimadas por el árbol para el pozo 3, resaltando que existen tres valores de permeabilidad calculados que son menores a cero.

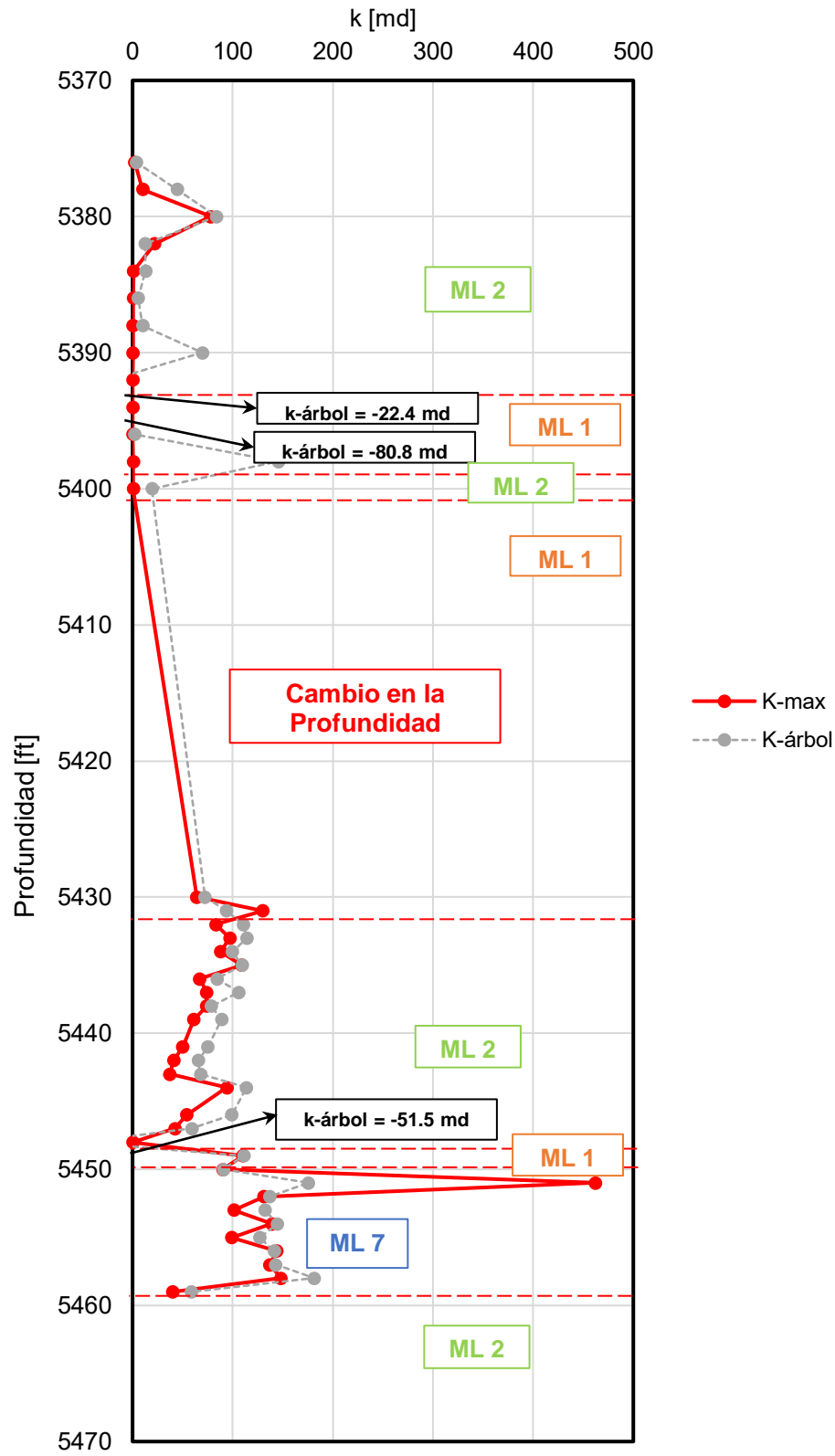


Figura 4.37. Permeabilidad Medida Contra Evaluada para el Pozo 3. (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).

Estos valores negativos fueron calculados a partir del modelo lineal 1 (ML1). Se deduce que los atributos con mayor peso en el ML (densidad de grano y profundidad) conducen al valor anómalo y que el registro GR y el ILD no son suficientes para explicar el fenómeno adecuadamente. Además, observando el comportamiento de todos los atributos en los tres puntos, especialmente el de la porosidad, se infiere que algo podría existir una obstrucción del medio poroso.

En el pozo 4 la **Figura 4.38** muestra el camino que recorren las instancias dentro del árbol hasta que alcanzan un nodo terminal.

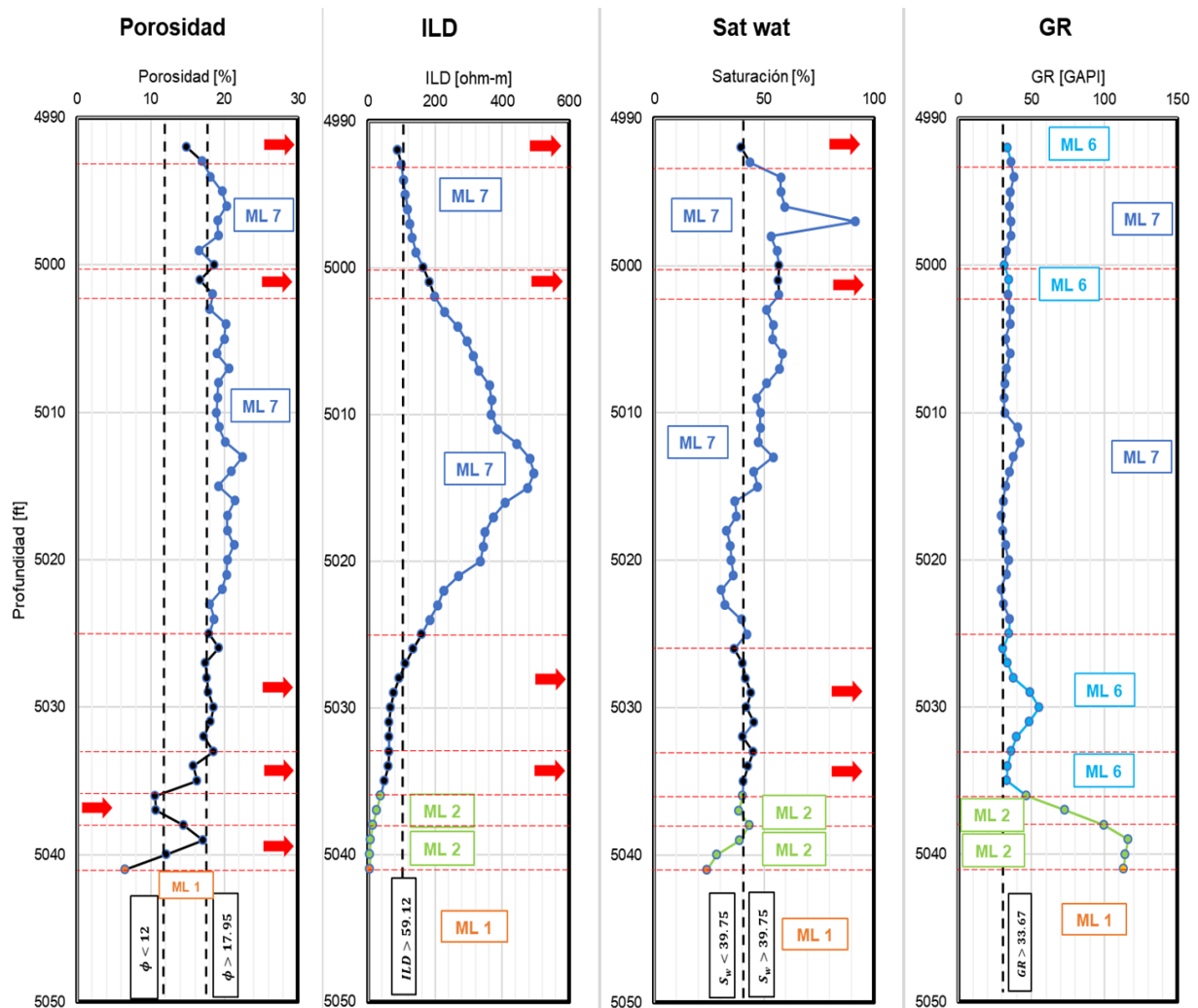


Figura 4.38. Procedimiento para el Pozo 4.

En la **Figura 4.39** se muestran los resultados obtenidos para este pozo comparando con aquellas mediciones de permeabilidad obtenidas de un análisis de núcleos en laboratorio,

dicha gráfica tiene las mismas características de formato que las presentadas para los pozos anteriores.

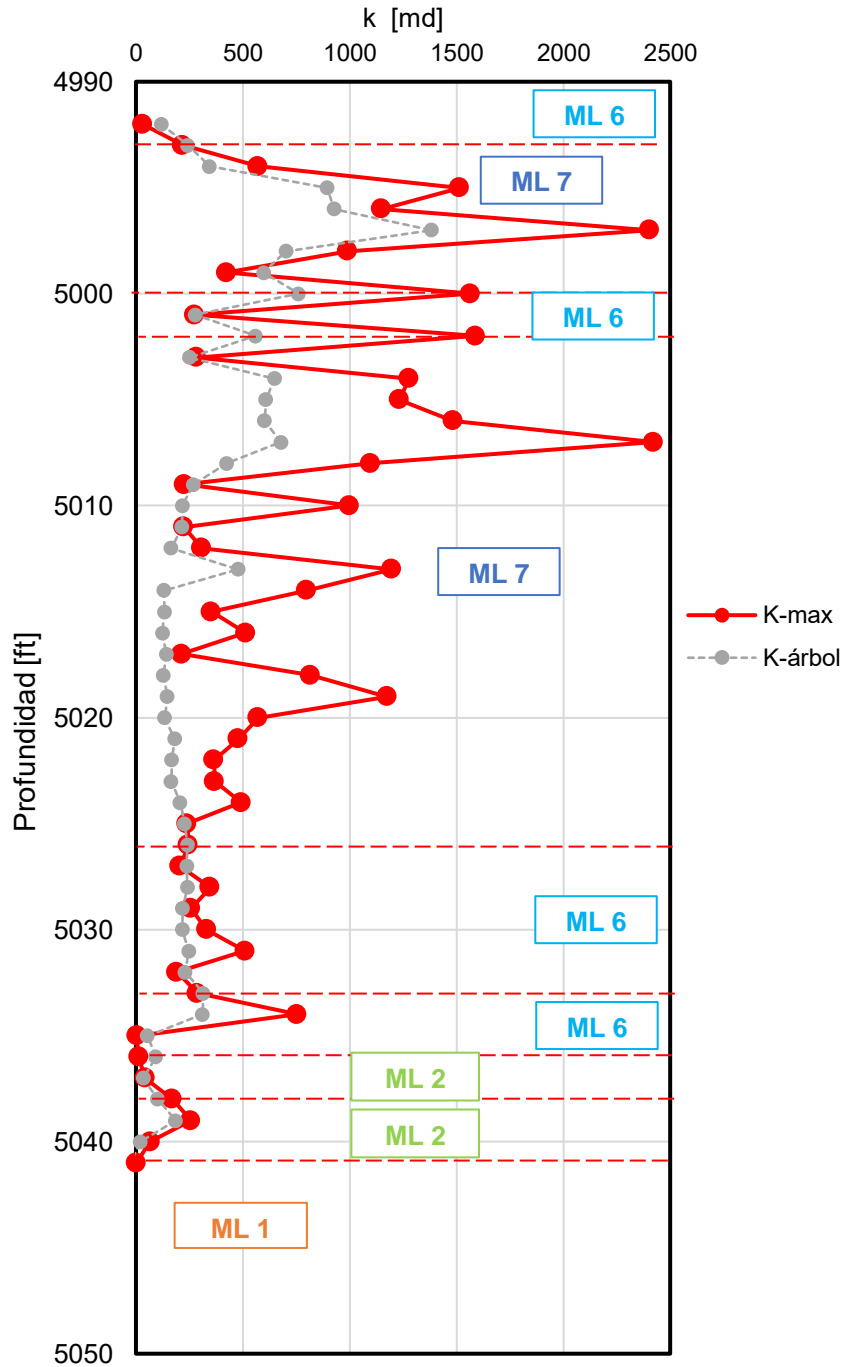


Figura 4.39. Permeabilidad Medida Contra Evaluada para el Pozo 4 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).

Para el pozo 5, **Figura 4.40**, se observan dos secciones de las instancias que pertenecen al modelo lineal 1 (*ML1*), en color naranja, después secciones con porosidades menores a 12% y una última entre 12% y 17.95% (color negro) que pasa a la evaluación con registro *ILD* y modelo lineal 2 (verde). Las estimaciones en la **Figura 4.41**

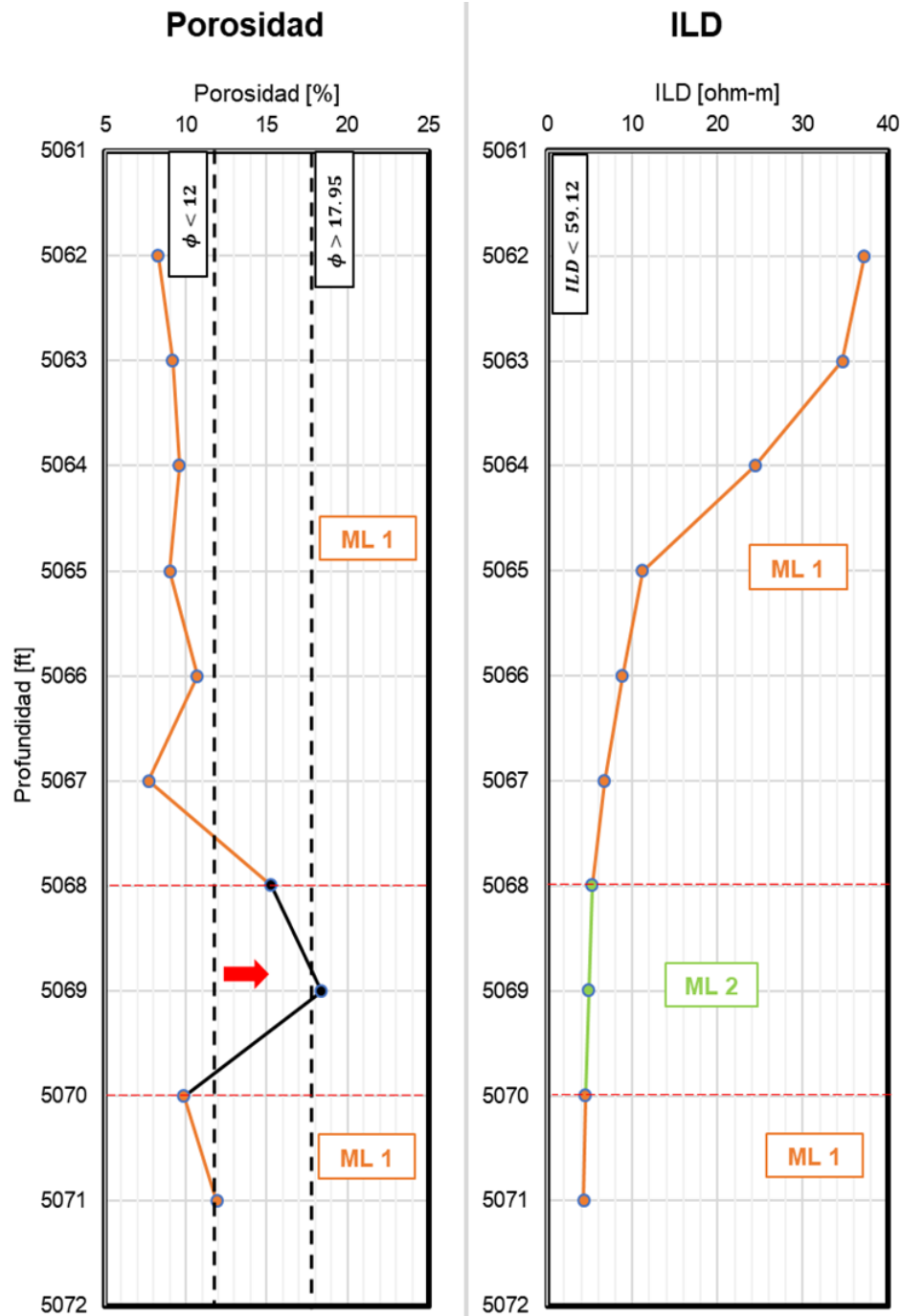


Figura 4.40. Procedimiento para el Pozo 5.

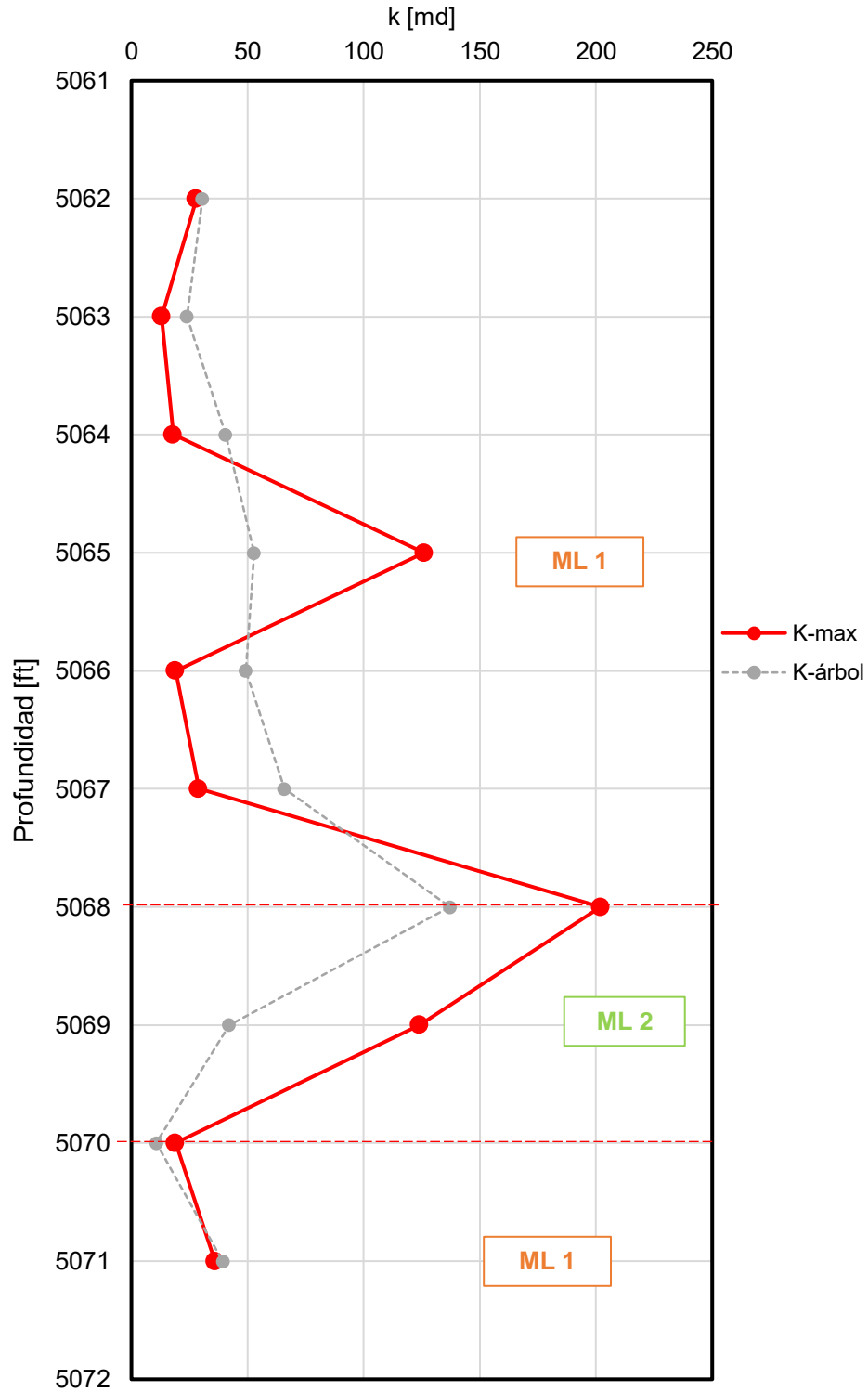


Figura 4.41. Permeabilidad Medida Contra Evaluada para el Pozo 5 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).

El procedimiento para el pozo 6 se muestra en la **Figura 4.42**, donde se observa que 4 secciones del total de instancias tienen porosidades mayores a 17.95% por lo que del nodo raíz pasan al nodo terminal del modelo lineal 7 (ML7) y las demás instancias son evaluadas con el modelo lineal 1 (ML1). En color negro se señalan las instancias que usan gráfico ILD y se estima con el modelo lineal 2 (ML2) y posteriormente con la saturación de agua y profundidad, modelo lineal 4 (ML4).

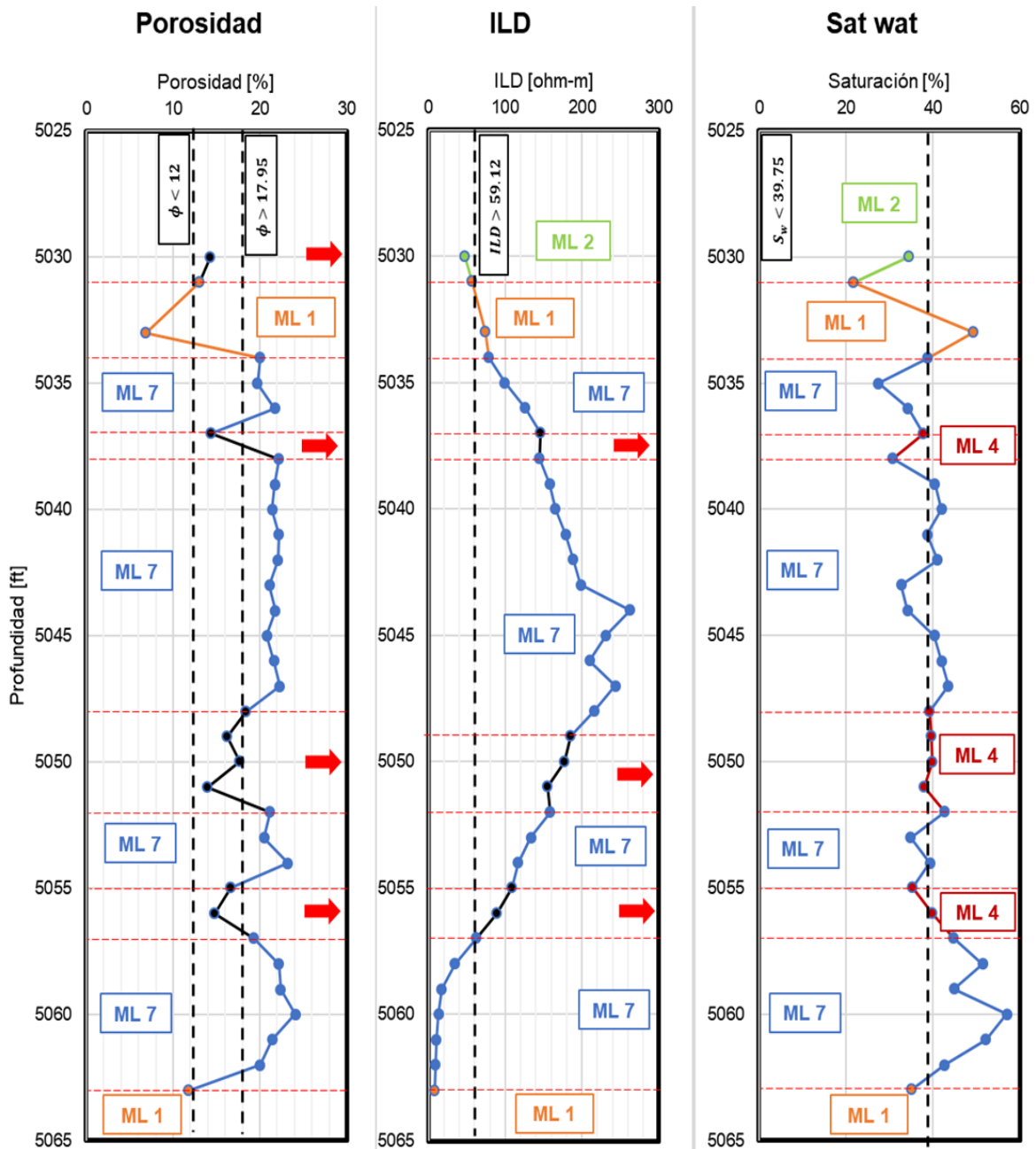


Figura 4.42. Procedimiento para el Pozo 6.

Los resultados para el pozo 6 se presentan en la **Figura 4.43**, donde se puede observar la existencia de un valor calculado negativo.

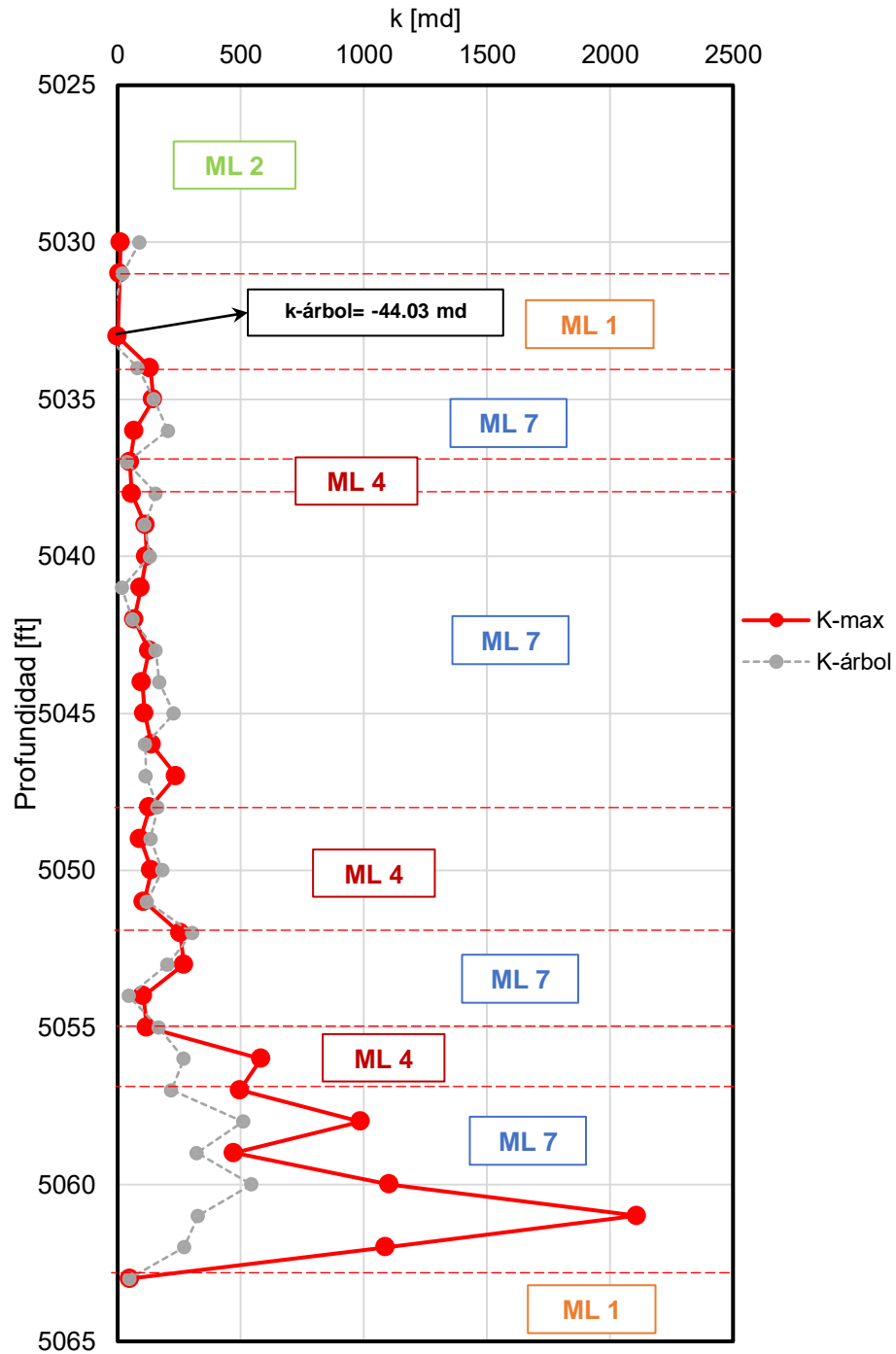


Figura 4.43. Permeabilidad Medida Contra Evaluada para el Pozo 6 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).

En la **Figura 4.44** se presenta el recorrido para el pozo 7 y, finalmente, en la **Figura 4.45** los resultados obtenidos.

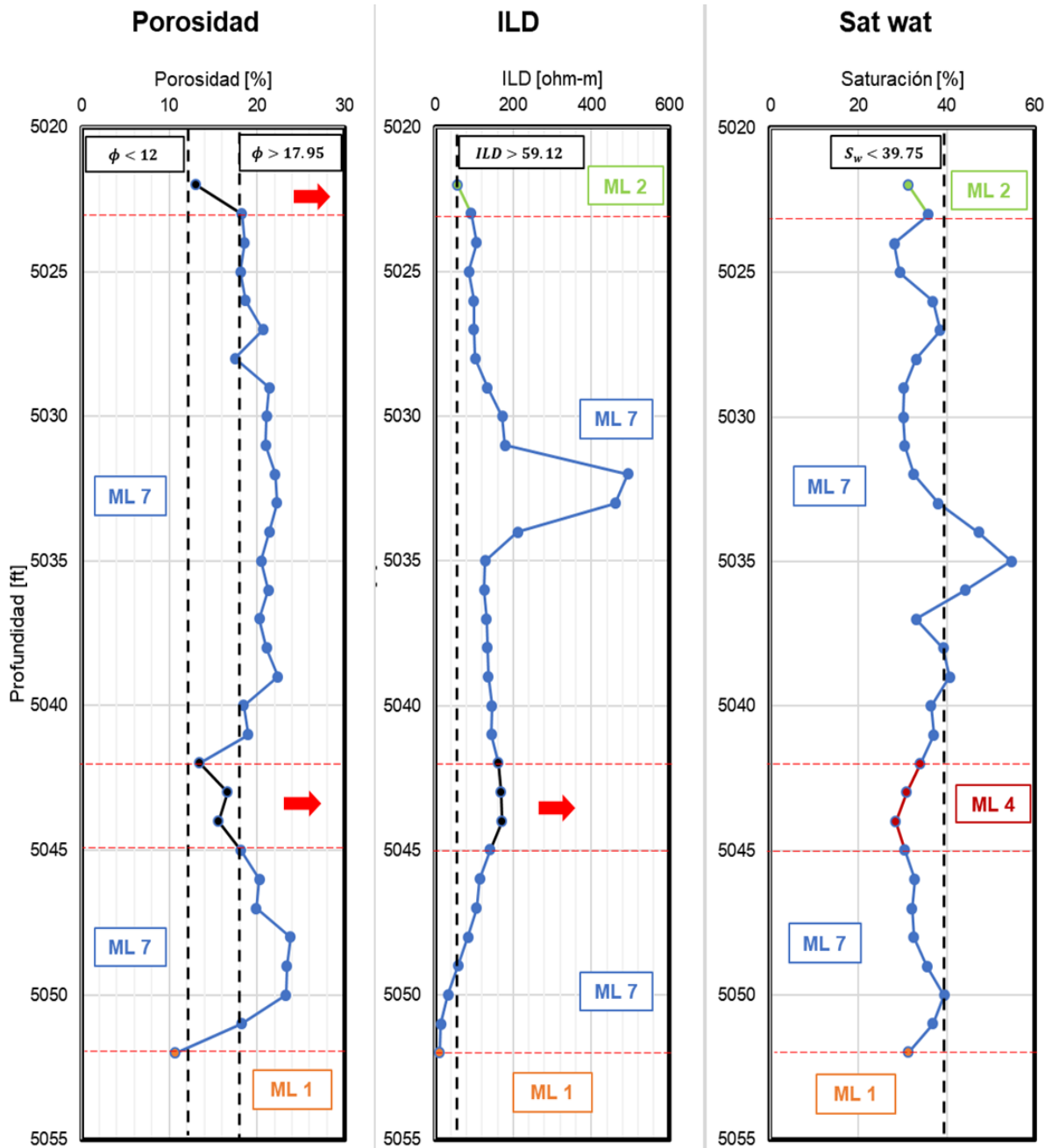


Figura 4.44. Procedimiento para el Pozo 6.

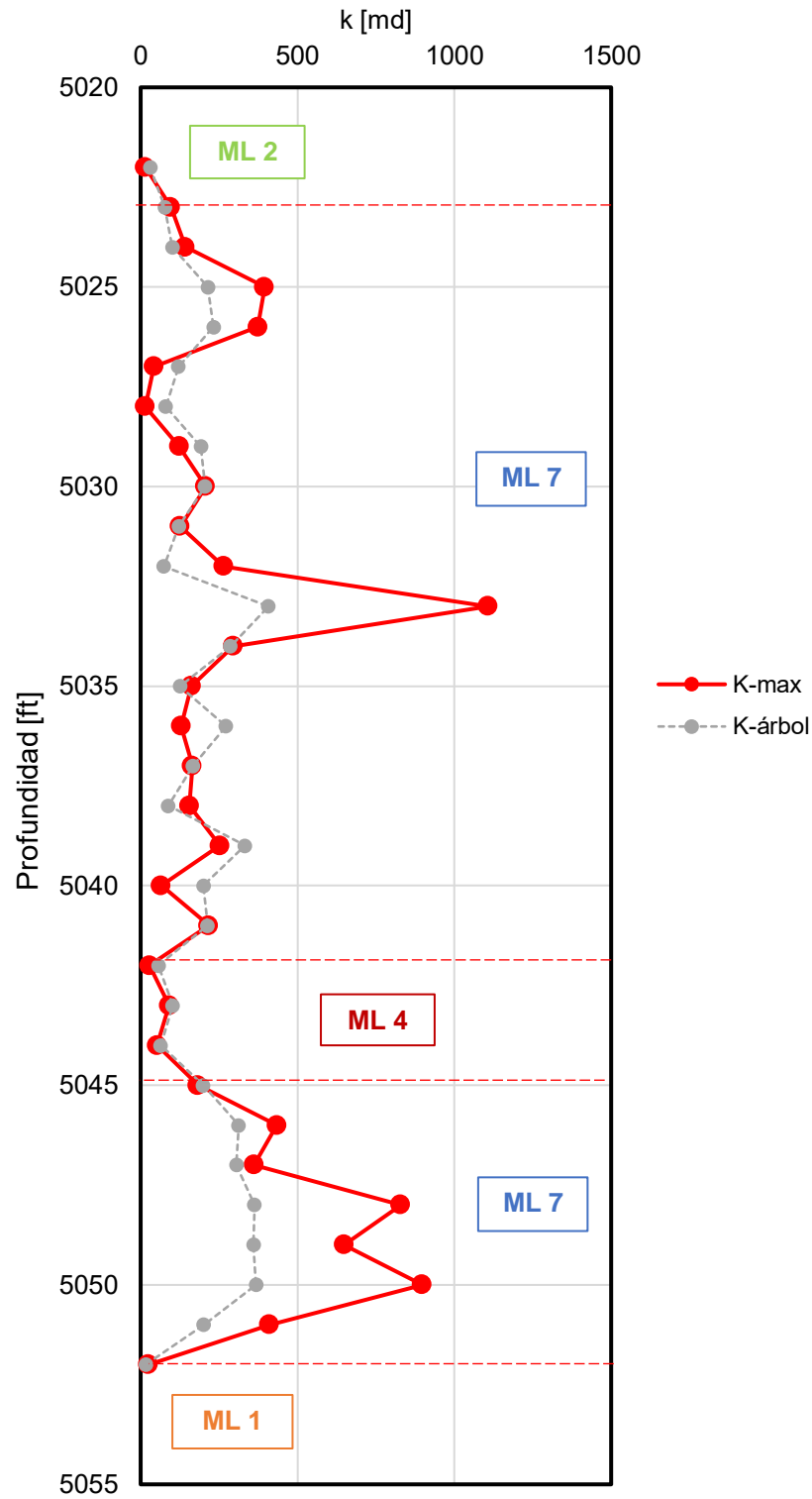


Figura 4.45. Permeabilidad Medida Contra Evaluada para el Pozo 7 (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).

La comparación de la permeabilidad medida contra evaluada para los siete pozos se presenta en la **Figura 4.46**. El árbol subestima en negativo sólo cuatro de 193 instancias. Los valores de R y R^2 son de 0.839 y 0.7044 respectivamente.

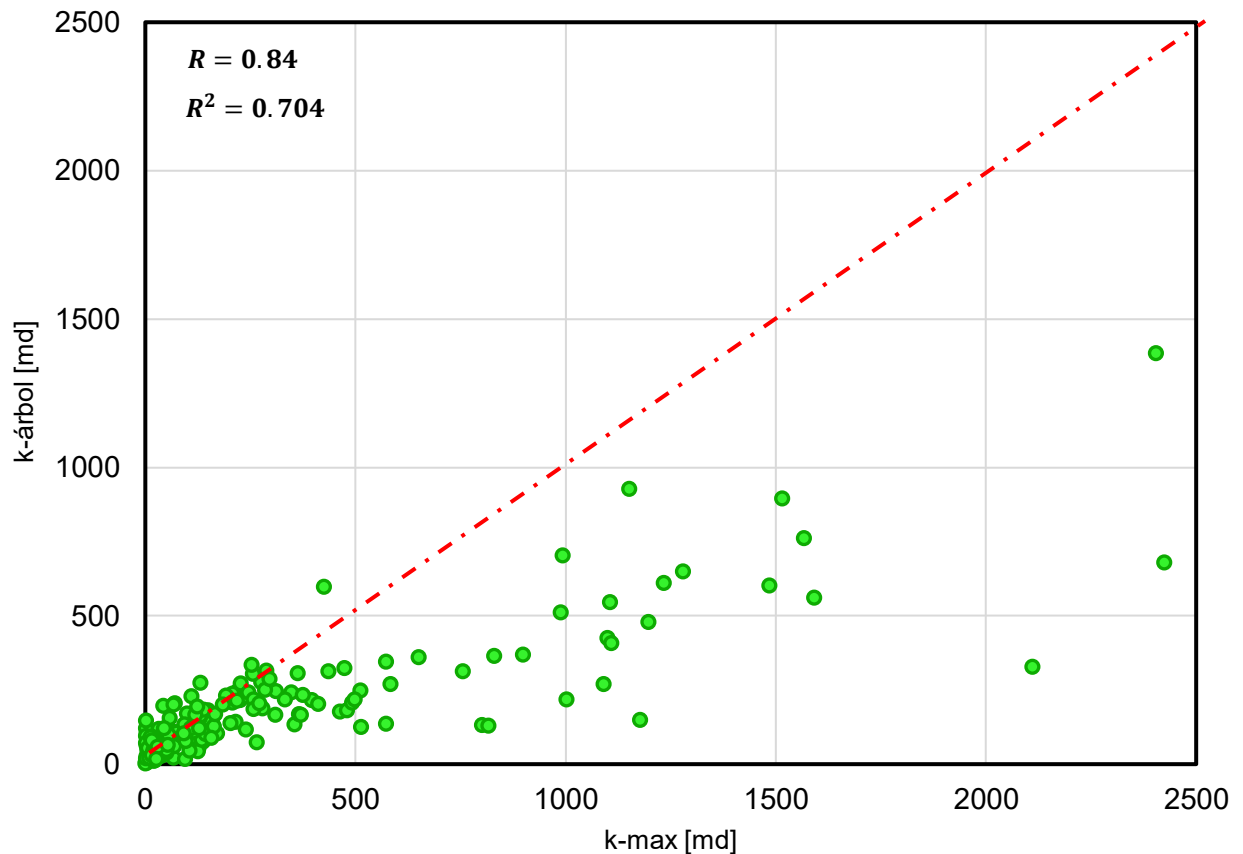


Figura 4.46. Gráfico de Valores de Permeabilidad Medidos contra Evaluados para los Siete Pozos (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).

El proceso sobre el pozo de validación se muestra en la **Figura 4.47**. En la **Figura 4.48** se observan los valores de permeabilidad estimados a partir de los atributos de entrada. En el conjunto de casos presentados hay núcleos con permeabilidades medidas y otros vacíos, la tendencia es perseguida muy cercanamente por el CART.

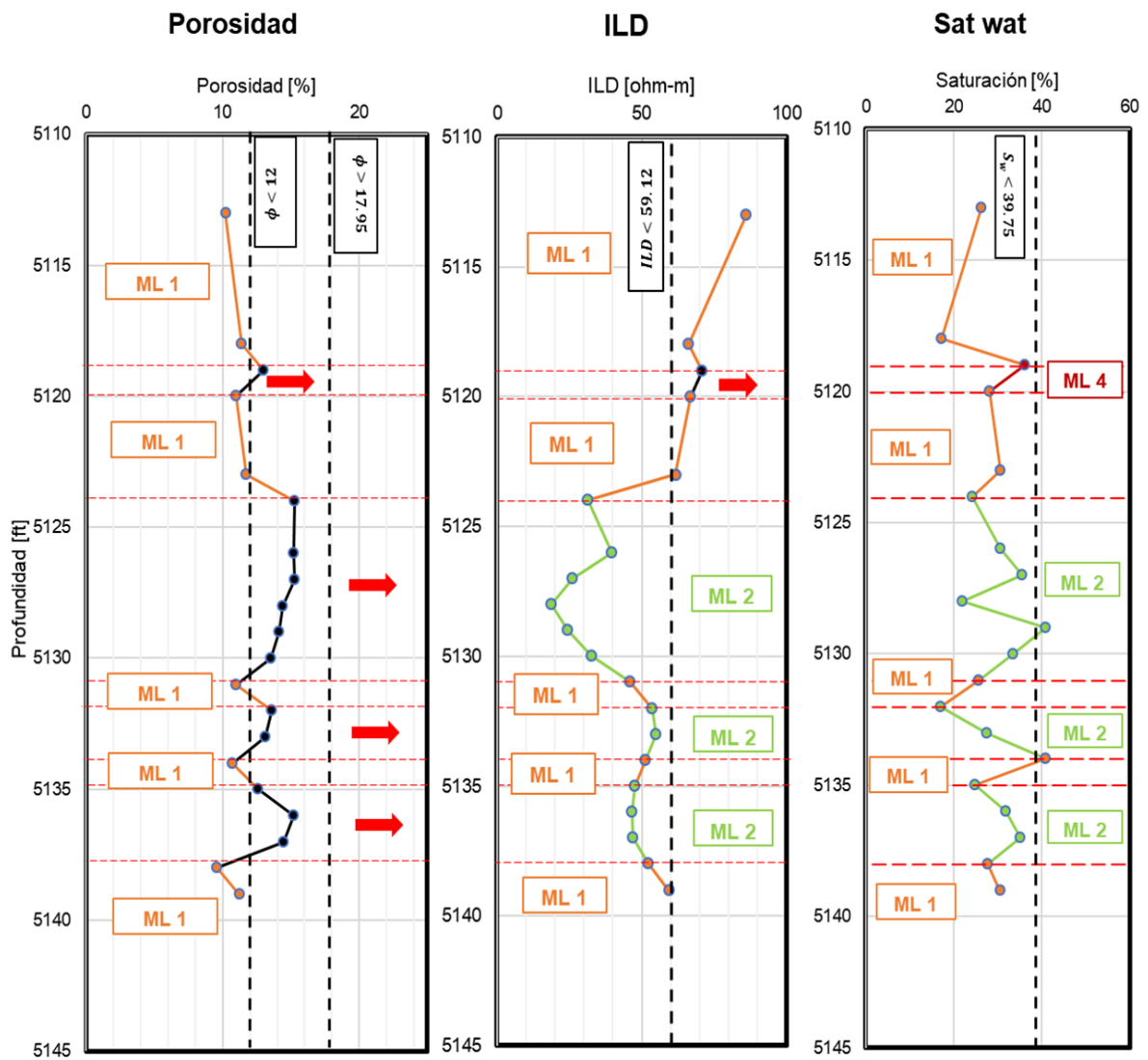


Figura 4.47. Procedimiento para el Pozo de Prueba.

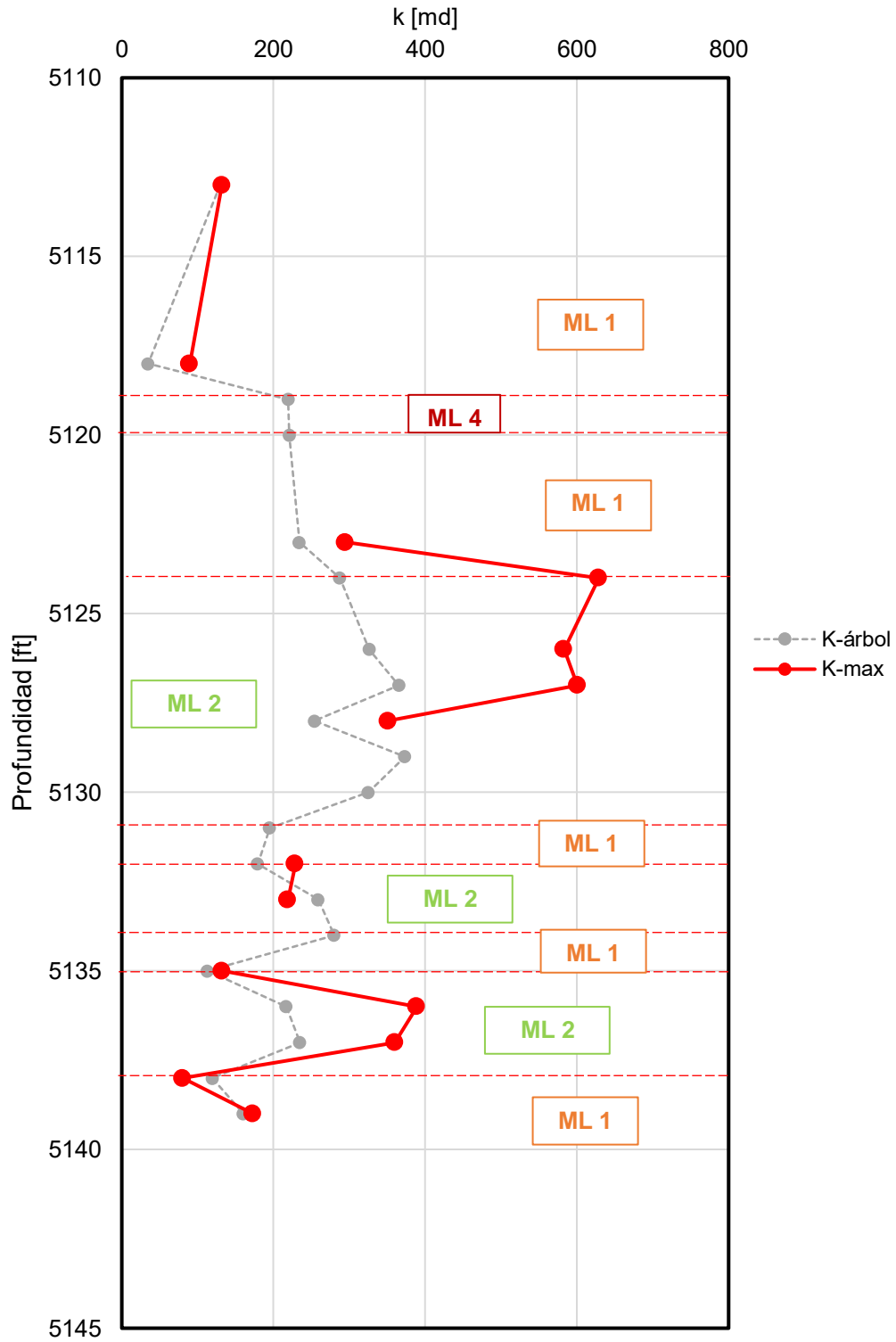


Figura 4 48. Estimación de los Valores de Permeabilidad para el Pozo de Prueba. (k-max: permeabilidad medida, k-árbol: permeabilidad estimada por el árbol).

4.4.2 Resultados: Análisis Factorial

Para decidir sobre los atributos fundamentales que explican al fenómeno en estudio, se determina la correlación o no-correlación entre estos y se reducen las dimensiones del problema. Primero es necesario asegurar que la base de datos es susceptible a la aplicación del análisis factorial con las pruebas de Esfericidad de Bartlett y la Medida de Adecuación Muestral Global (KMO).

En la prueba de Esfericidad de Bartlett se busca que la matriz de correlaciones sea lo suficientemente distinta de la matriz identidad con el p-valor menor a 0.05. Para el caso de medida KMO, el resultado debe ser mayor a 0.6. En la **Figura 4.49** se muestra la matriz de correlaciones de la base de datos y en la **Tabla 4.3** los valores de las pruebas Bartlett y KMO.

	DEPTH	ILD	GR	NPHI	Porosidad	Satoil	Satwat	Dengran	Kmax	
Correlación	DEPTH	1.000	-.296	.155	-.094	.275	-.019	-.336	-.787	.011
	ILD	-.296	1.000	-.359	.032	.385	.311	-.156	-.018	.367
	GR	.155	-.359	1.000	.487	-.149	.104	-.053	.002	.030
	NPHI	-.094	.032	.487	1.000	.174	.690	-.414	.023	-.047
	Porosidad	.275	.385	-.149	.174	1.000	.449	-.474	-.586	.444
	Satoil	-.019	.311	.104	.690	.449	1.000	-.730	-.116	-.013
	Satwat	-.336	-.156	-.053	-.414	-.474	-.730	1.000	.390	.047
	Dengran	-.787	-.018	.002	.023	-.586	-.116	.390	1.000	-.277
	Kmax	.011	.367	.030	-.047	.444	-.013	.047	-.277	1.000

Figura 4.49. Matriz de Correlaciones.

Se puede observar que en la matriz de correlaciones existe relación entre variables y a primera vista pareciera ser lo suficientemente distinta de una matriz identidad, lo que se verifica al ver los resultados de la tabla 4.3: prueba de esfericidad < 0.05.

Tabla 4.3. Resultados de las pruebas de Esfericidad de Bartlett y KMO.

Medida KMO de Adecuación de Muestreo	0.603	
Prueba de Esfericidad de Bartlett	Aprox. Chi-cuadrado	988.293
	Grados de Libertad	36
	Significancia	0.000

Por otra parte, la matriz de significancia de la **Figura 4.50** muestra que existe relación entre las variables. La prueba KMO al mayor a 0.6 confirma que existen fuertes correlaciones entre grupos de variables y correlaciones pequeñas entre pares de variables. Entonces es posible hacer uso del análisis factorial.

Sig. (unilateral)	DEPTH		.000	.016	.096	.000	.394	.000	.000	.437
	ILD	.000		.000	.331	.000	.000	.015	.401	.000
	GR	.016	.000		.000	.019	.075	.233	.487	.340
	NPHI	.096	.331	.000		.008	.000	.000	.373	.259
	Porosidad	.000	.000	.019	.008		.000	.000	.000	.000
	Satoil	.394	.000	.075	.000	.000		.000	.055	.428
	Satwat	.000	.015	.233	.000	.000	.000		.000	.256
	Dengran	.000	.401	.487	.373	.000	.055	.000		.000
	Kmax	.437	.000	.340	.259	.000	.428	.256	.000	

Figura 4.50. Matriz de Significancia.

En el gráfico de sedimentación (a partir de la regla de Kaiser) (**Figura 4.51**) se determinó que el número óptimo de factores con los que se puede representar el fenómeno estudiado es cuatro. Esta regla indica que los factores a escoger son todos aquellos con autovalores mayores a 1. Es importante recordar que un autovalor es una medida de la varianza que explica cuánta información puede ser explicada por un factor.

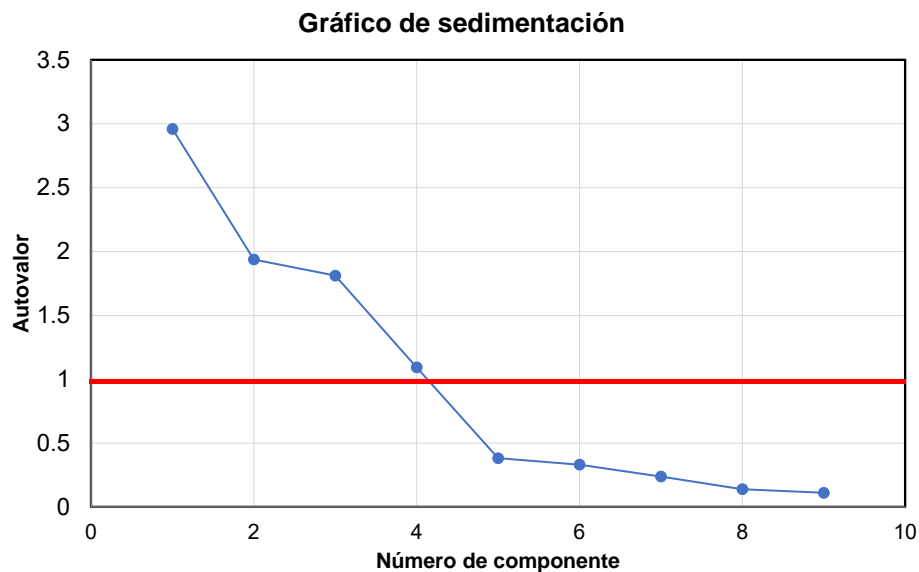


Figura 4.51. Gráfico de Sedimentación.

En la **Tabla 4.4** se presentan las comunalidades del modelo factorial, es decir, la proporción de varianza que éste es capaz de reproducir por cada atributo de forma individual.

Tabla 4 4. Comunalidades del Modelo Factorial.

Comunalidades	
Profundidad	0.912
Registro ILD	0.784
Registro GR	0.898
Registro NPHI	0.865
Porosidad	0.795
Saturación de Aceite	0.914
Saturación de Agua	0.830
Densidad de Grano	0.902
Permeabilidad	0.900

Como se puede observar, el modelo factorial puede explicar, para todos los atributos, arriba del 75% de la varianza, donde el atributo mejor representado es la Saturación de Aceite con una explicación de su varianza del 91%, mientras que aquel con la menor varianza explicada es el Registro ILD con 78%.

En la **Tabla 4.5** se expone la varianza total explicada por factor. Esto muestra la varianza que puede explicar cada factor, así como la parte de la varianza explicada por el modelo factorial de la total que representa el fenómeno.

Tabla 4.5. Varianza Total Explicada.

Componente	Sumas de cargas al cuadrado de la rotación		
	Total	% de varianza	% Acumulado
1	2.475	27.496	27.496
2	2.168	24.089	51.585
3	1.661	18.460	70.045
4	1.496	16.621	86.666

El modelo factorial puede explicar el 86.66% de la varianza total de la base de datos, y debido a que el método de extracción usado fue el de análisis de componentes principales, el primer factor determinado es aquél que puede describir la mayor parte de la variación en los datos, en este caso el 27.496%, donde para los factores extraídos subsecuentemente este valor va decreciendo hasta llegar al factor cuarto y último que es capaz de explicar el 16.62% de la varianza total.

La matriz de componentes rotados (**Figura 4.52**) señala donde se encuentran los factores formados con los respectivos atributos que los componen (con las magnitudes de la correlación entre el factor y la variable, conocidas como cargas factoriales).

	Componente			
	1	2	3	4
Satoil	.951			
Satwat	-.820			
NPHI	.740			.536
DEPTH		-.941		
Dengran		.905		
Kmax			.927	
ILD			.610	
Porosidad			.582	
GR				.942

Figura 4.52. Matriz de Componente Rotado.

El primer factor está conformado por los atributos Saturación de Aceite, Saturación de agua y por el Registro NPHI, el segundo por la Profundidad y la Densidad de Grano, el tercero por la Permeabilidad, el Registro ILD y la Porosidad y finalmente el cuarto componente tiene un atributo, el Registro GR.

Interpretación de los Factores

Con base en las similitudes conceptuales y teóricas de las variables que conforman a los factores se presenta su nombre y significado.

Factor 1 *Estado de la fase no sólida*

Este factor, **Figura 4.53**, está conformado por los atributos:

- Saturación de aceite.
- Saturación de agua → inversa.
- Registro NPHI.

Si la saturación es el volumen ocupado por un fluido en relación con el volumen poroso total disponible del medio y el registro NPHI mide la energía en un flujo de electrones que es disparado hacia la formación (mayor saturación de agua y/o aceite menor será la energía - mayor la medición), este factor agrupa características del medio sobre el estado relativo de la fase no sólida, en palabras simples: la condición de “llenado” en la que se encuentra el espacio poroso disponible.

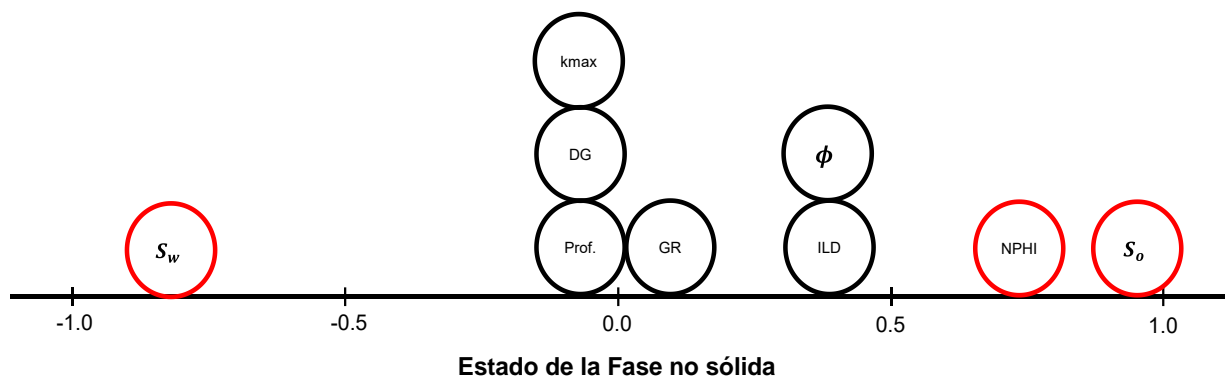


Figura 4.53. Grafico del componente 1 en espacio rotado.

La **Figura 4.54** muestra, relacionado con este factor, las instancias con mayores saturaciones de agua, de aceite y con mayores o menores mediciones del registro NPHI. Aquí se resaltan los atributos pertenecientes a un mismo pozo en cada una de las tres secciones.

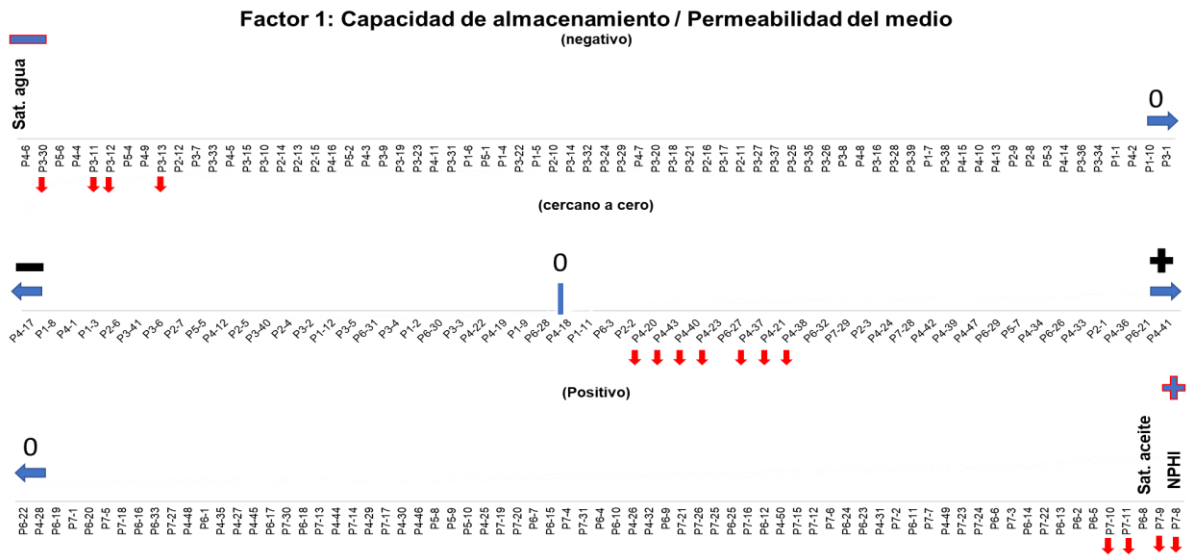


Figura 4.54. Instancias del Factor 1.

Las cuatro instancias resaltadas (con flechas) en la sección más negativa pertenecen al pozo 3 en la formación Morrow en el suroeste de Kansas, con mayor saturación de agua y la menor saturación de aceite y registro NPHI. Las instancias resaltadas en la sección más positiva del factor pertenecen al pozo 7 encontrado en la formación Morrow superior en el suroeste de Kansas y presentan una mayor saturación de aceite y mediciones del registro neutrón con las menores saturaciones de agua.

Factor 2 Sepultamiento.

Este segundo factor, **Figura 4.55**, está conformado por los atributos siguientes:

- Profundidad.
- Densidad del grano → inversa.

Entiéndase a la profundidad como la distancia que existe entre un plano horizontal de referencia y la distancia vertical negativa a la que se midieron los atributos. La densidad de grano (o densidad matricial) es la relación entre la masa y el volumen de la matriz, sin considerar el volumen del espacio poroso. La relación inversa deriva del sesgo por el tamaño de la matriz de datos. Debe señalarse que existen, por ejemplo, secciones de arcillas consolidadas y de calizas con densidades altas en secciones menos profundas lo que genera este tipo de resultados.

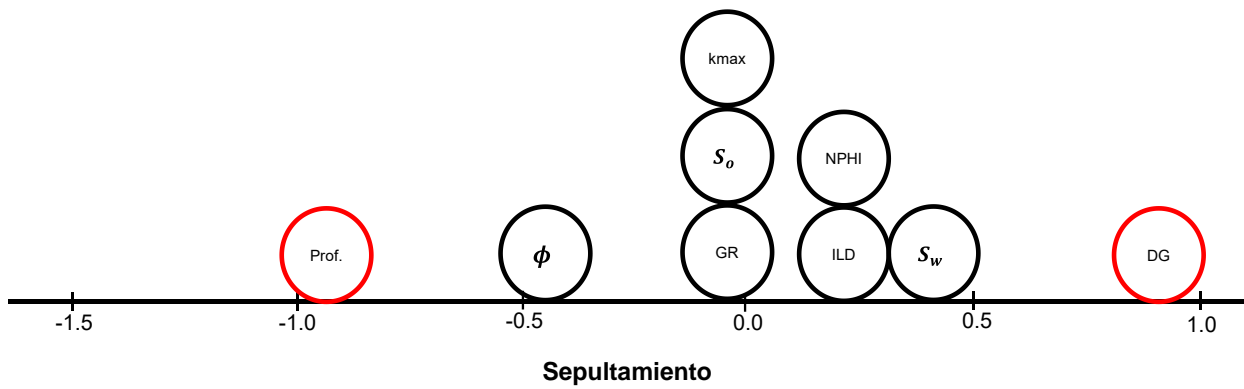


Figura 4.55. Gráfico del Componente 2 en Espacio Rotado.

En la **Figura 4.56** se presentan el segundo factor y algunas instancias. Aquellas encontradas en la parte más negativa del factor pertenecen al pozo 3, ubicado en el suroeste de Kansas en la formación Morrow (inferior y media) con las menores densidades (areniscas de grano muy fino a fino con poca presencia de calcita). Las instancias resaltadas del lado positivo pertenecen al pozo 2, ubicado en el noreste de Kansas en la formación Viola (dolomitas), con las mayores densidades entre todas las muestreadas. Las instancias cercanas a cero son del pozo 6 de la formación Morrow (areniscas lenticulares) que son las que tienen tanto una densidad y profundidad cercana a la mediana (areniscas de grano grueso con buenas porosidades y poca presencia de otras litologías).

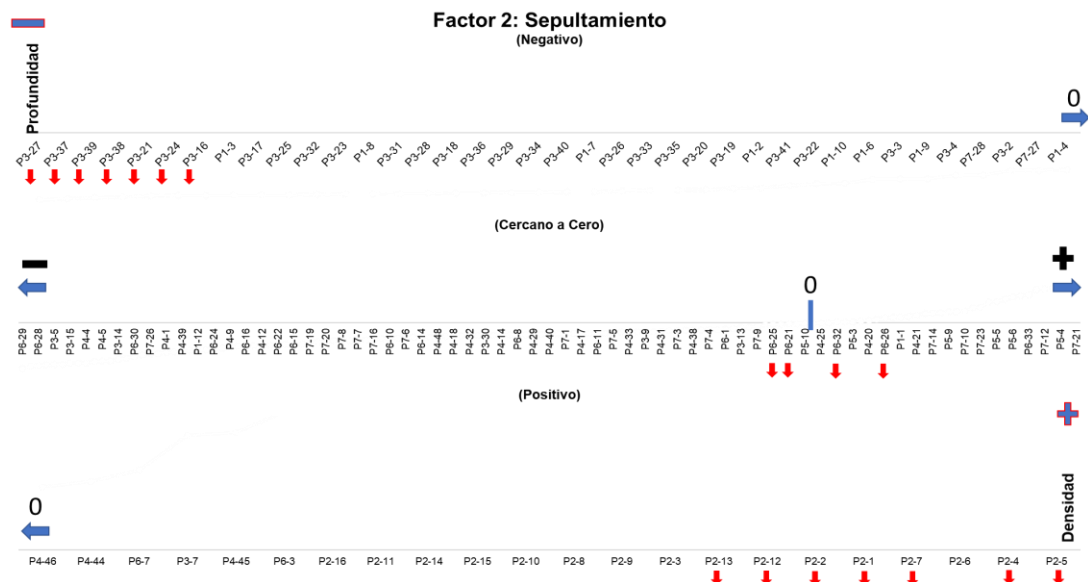


Figura 4.56. Instancias del Factor 2.

Factor 3 Estado de la fase sólida

Este tercer factor, **Figura 4.57**, está conformado por los siguientes atributos:

- Permeabilidad.
- Registro ILD.
- Porosidad.

Si la permeabilidad es la capacidad de una roca para dejar pasar fluido a través de ésta (liga obvia con la porosidad) y el registro de inducción o resistividad ILD mide la capacidad de impedir el paso de una corriente eléctrica (agua salada líquido altamente conductor, aceite fluido altamente resistivo) entonces se propone que la relación entre los tres atributos significa en la “forma” del medio sólido como medio para conducir cualquier fluido.

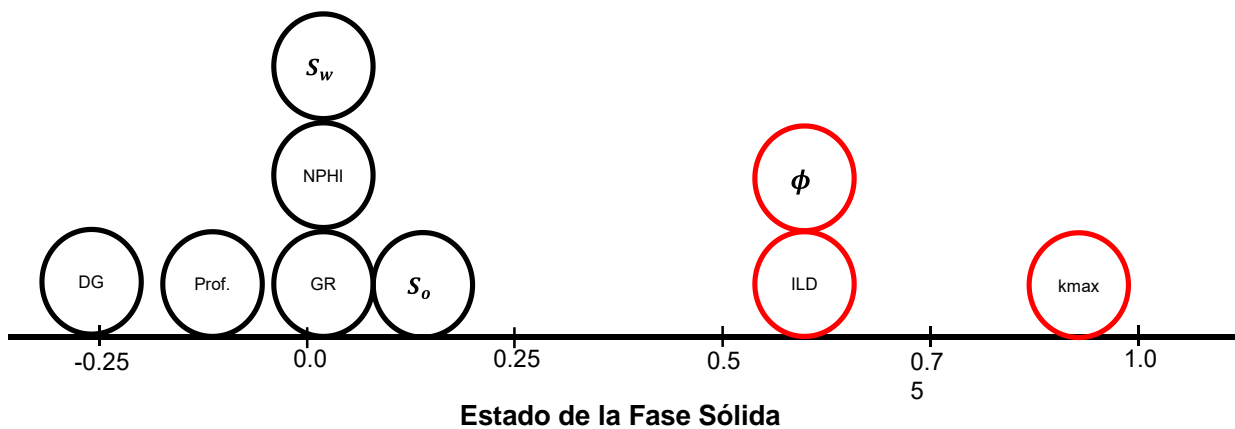


Figura 4.57. Gráfico del Componente 3 en Espacio Rotado.

En este importante tercer factor (explica por sí solo la capacidad del medio para ser productor) (**Figura 4.58**) las instancias resaltadas en la sección más negativa (menor porosidad y permeabilidad) son algunas que pertenecen al pozo 3. En el lado positivo se muestran casos del pozo 4, al suroeste de Kansas en la formación Morrow superior (pozo de gas).

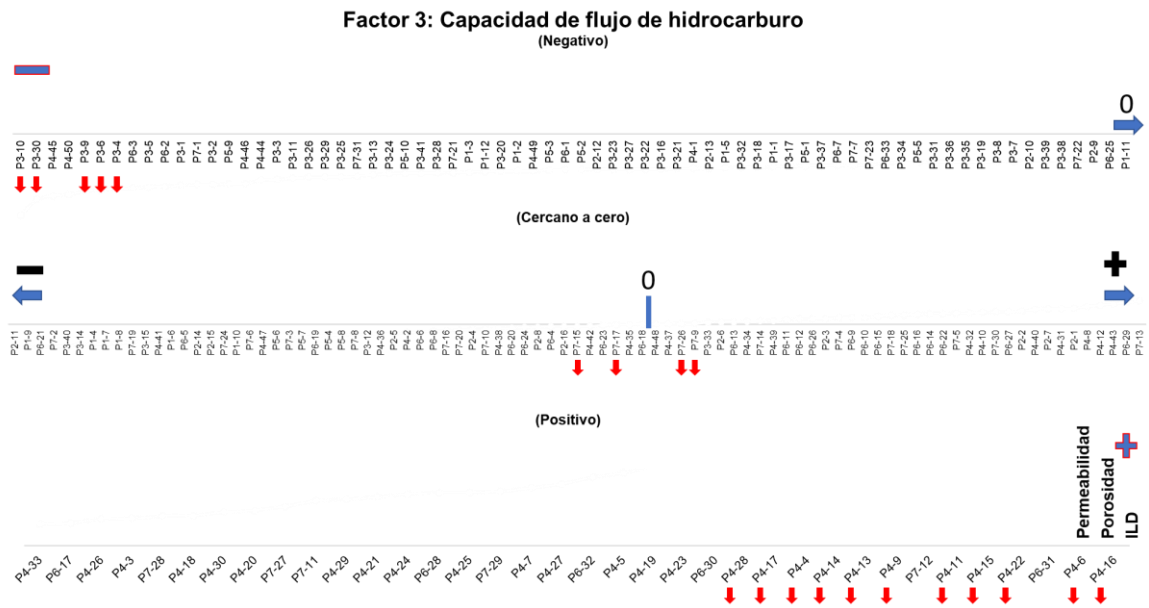


Figura 4.58. Instancias del Factor 3.

Factor 4

El último factor, **Figura 4.59** cuenta con un atributo:

- Registro GR

Derivado de la radiactividad emitida se define, por ejemplo, la cantidad de arcillas presentes en una formación y, por lo tanto, una potencial reducción de la permeabilidad. Esta propiedad se aparta de las condiciones anteriores que son relacionables con aspectos físicos de un material sólido que permite el flujo de sustancias por la disposición efectiva de sus espacios vacíos. Por esta razón no se le asigna nombre. En la **Figura 4.60** se muestran las instancias en la regla de variación.

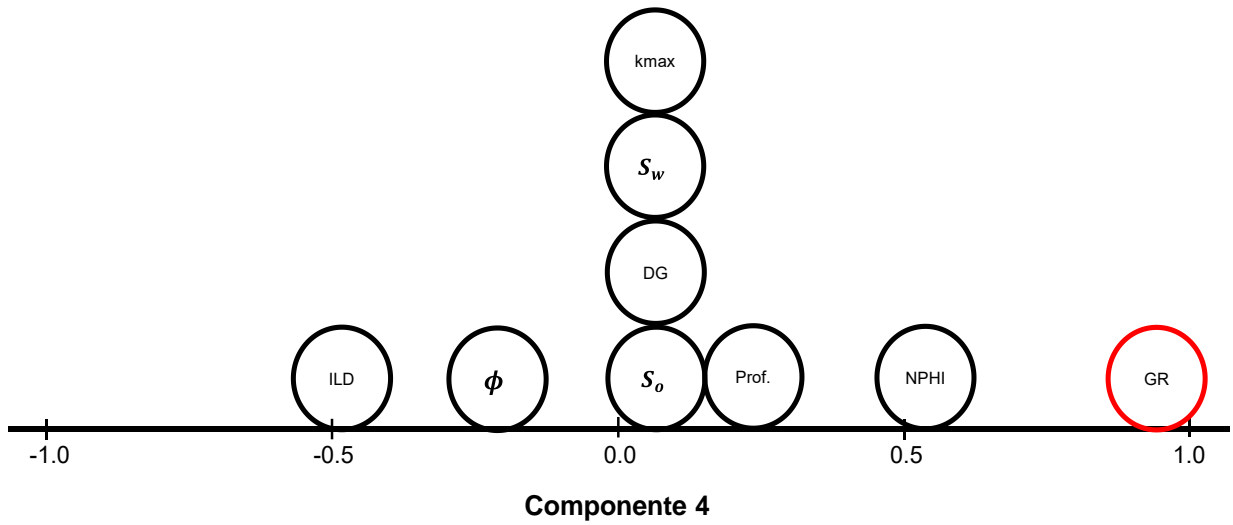


Figura 4.59. Gráfico del Componente 4 en Espacio Rotado.

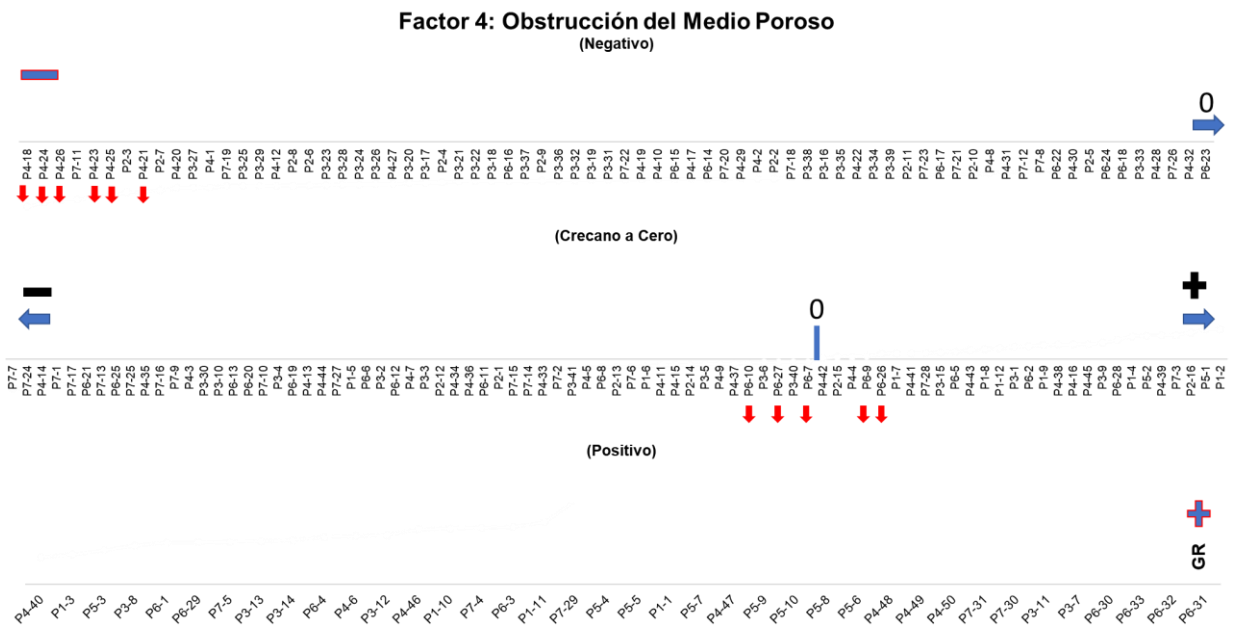


Figura 4.60. Instancias Factor 4.

4.4.3 Resultados: Redes Neuronales.

Se presenta el proceso para obtener un modelo neuronal para estimar la permeabilidad. Se resaltan las arquitecturas probadas y las singularidades de su entrenamiento y prueba.

El modelo consta de ocho nodos de entrada:

- profundidad de la muestra,
- registro de resistividad (ILD),
- registro de rayos gamma (GR),
- registro neutrón porosidad (NPHI),
- porosidad medida de la muestra,
- saturación de agua,
- saturación de aceite y
- densidad de grano.

Todos los modelos tienen un nodo de salida:

- permeabilidad.

En el entrenamiento las salidas de la red se comparan con la permeabilidad objetivo (entrenamiento supervisado) de la base de datos. Usando el algoritmo de aprendizaje *backpropagation* se actualizan los pesos del sistema de acuerdo con la definición de error. Este proceso (iterativo) continúa hasta que se alcanza la diferencia acordada como aceptable entre lo medido y lo estimado. En la etapa de entrenamiento se usó el algoritmo de validación cruzada (k-folds cross-validation) para mejorar los resultados obtenidos, tomando el 65% de las instancias para entrenamiento (124 instancias), el 20% para prueba (41 instancias) y el 15% (28) para ser usadas como validación (selección aleatoria). La función de error para todos los modelos neuronales es la suma de cuadrados.

Modelo 1 (tangente hip-tangente hip)

Este primer modelo consiste en ocho nodos de entrada, una capa oculta (se probaron en una capa con 10 y hasta 270 unidades) y un nodo de salida (permeabilidad). La función

de activación de la capa oculta y de salida fue tangente hiperbólica. La tasa de aprendizaje fue 0.57, 250000 iteraciones.

La **Figura 4.61** muestra el comportamiento del coeficiente de correlación R para los distintos arreglos de nodos ocultos. Las barras color naranja corresponden al coeficiente R en entrenamiento y las barras amarillas al coeficiente obtenido en la prueba. El modelo con 270 nodos en la capa oculta es el mejor.

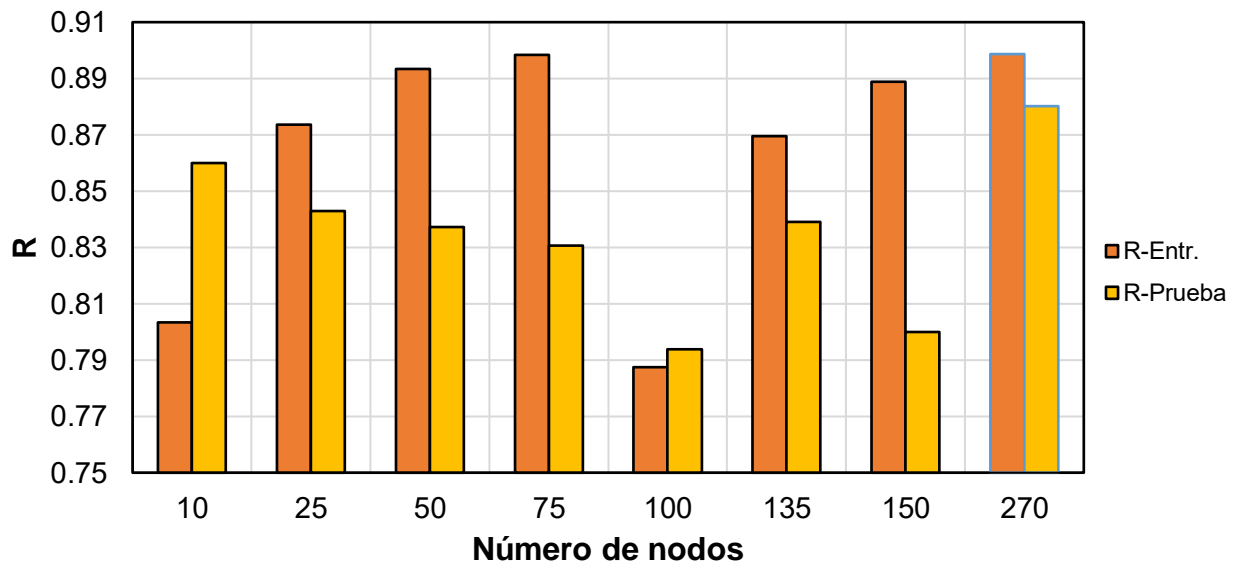


Figura 4.61. Calificación de nodos para el modelo 1, (R-Entr: Coeficiente de correlación del entrenamiento, R-Prueba: Coeficiente de correlación de la prueba).

En la **Figura 4.62** se muestra el resumen de las características del modelo 1.

Entrenamiento	Error de suma de cuadrados	1.492
	Error relativo	.159
	Regla de parada utilizada	1215000 paso(s) consecutivo(s) sin disminución del error ^a
	Tiempo de entrenamiento	0:27:40.81
Pruebas	Error de suma de cuadrados	.686
	Error relativo	.390
Reserva	Error relativo	.365

Figura 4.62. Resumen del Modelo de Red 1.

La importancia relativa de los atributos de entrada en el modelo neuronal se muestra en la **Figura 4.63**. Este resultado debe ser analizado con precaución ya que de acuerdo con lo mostrado en los análisis de CD el factor que involucra a la profundidad tiene un sesgo que podría no responder al fenómeno analizado. La saturación de agua y porosidad resultan estar más relacionadas con la permeabilidad y, aunque el modelo parece prometedor, se intentan otros arreglos por la poca sonoridad teórica.

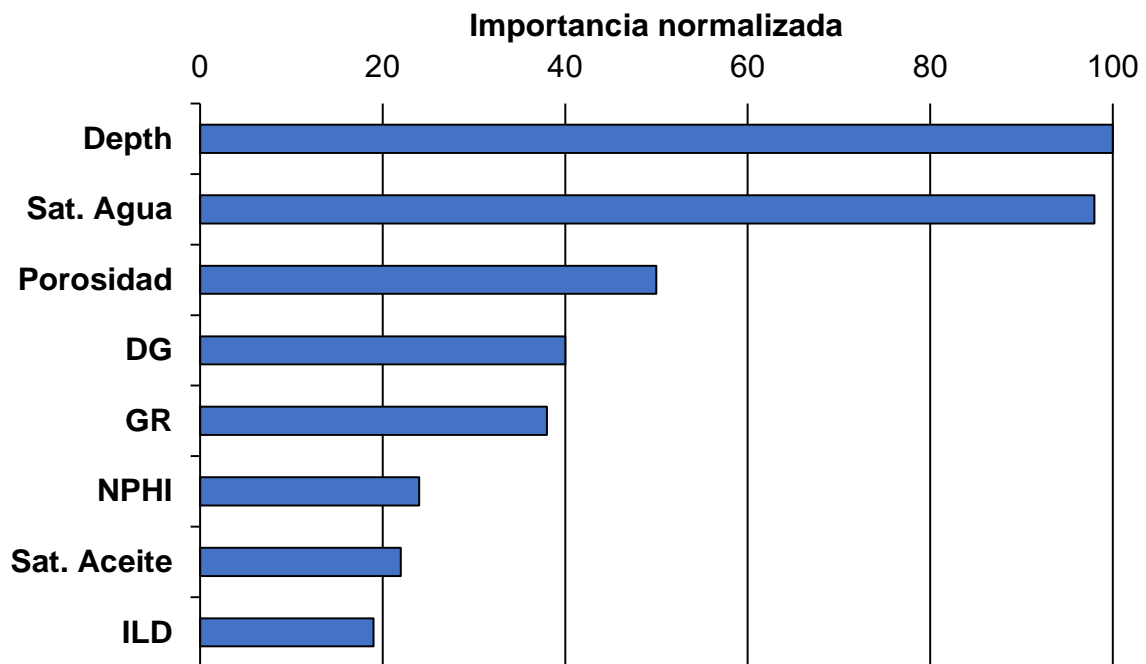


Figura 4.63. Importancia de los Atributos de Entrada para el Modelo de Red 1.

En la **Figura 4.64** se observan los valores medidos contra evaluados por este modelo neuronal. Se puede apreciar que esta red subestima los valores más altos de permeabilidad y se distinguen errores importantes en las permeabilidades muy pequeñas.

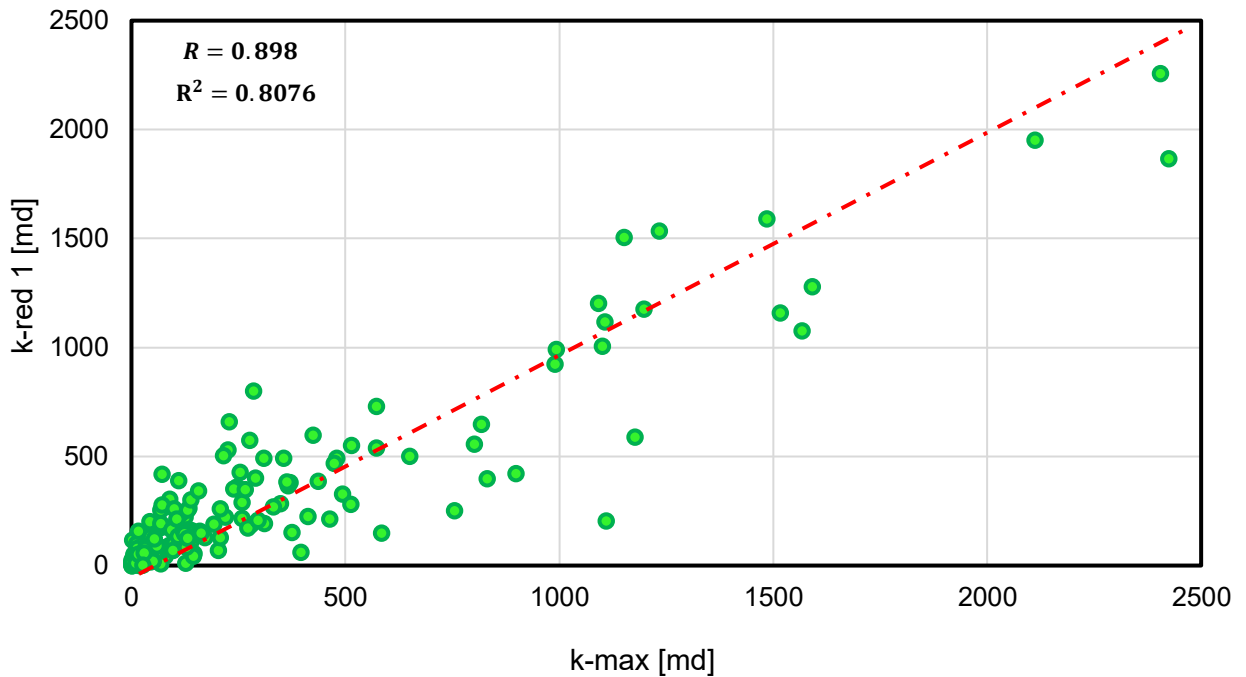


Figura 4.64. Permeabilidad Medida Contra Evaluada para la Red 1 (8 entradas, 270 nodos, 1 salida).

La comparación medidos contra evaluados en prueba se presenta en la **Figura 4.65**.

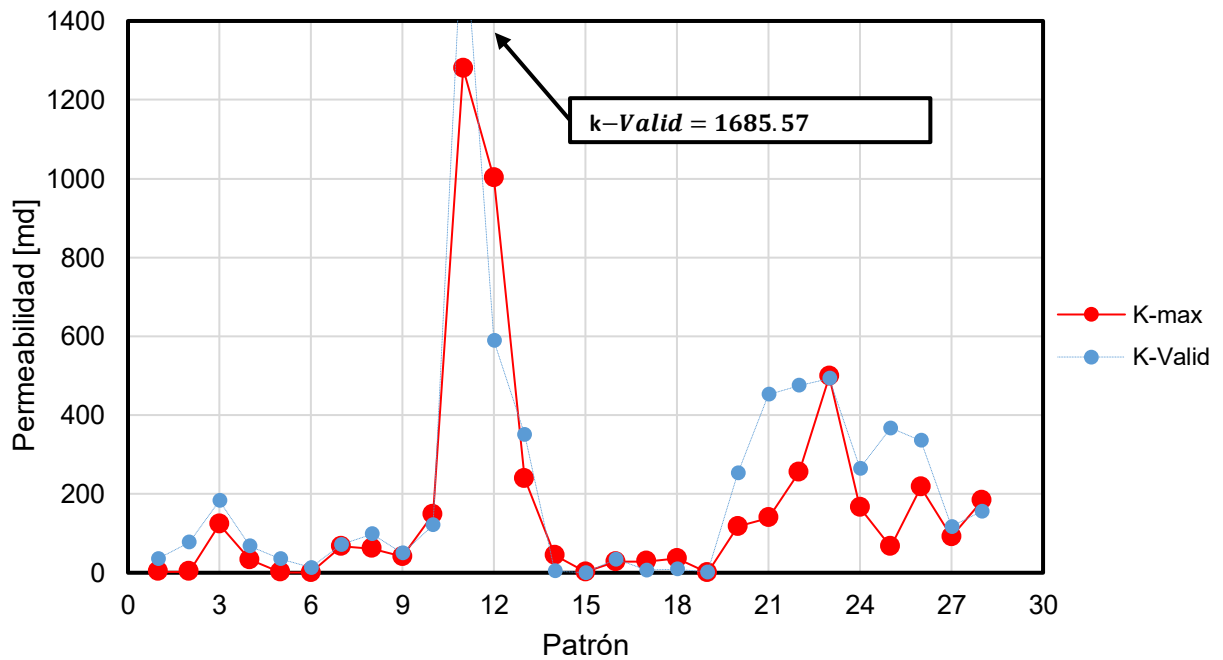


Figura 4.65. Predicción de Permeabilidad por la Red 1 (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).

Finalmente, en la **Figura 4.66**, se exhibe la aplicación sobre el pozo de prueba, permeabilidades desconocidas en algunos de los puntos muestreados.

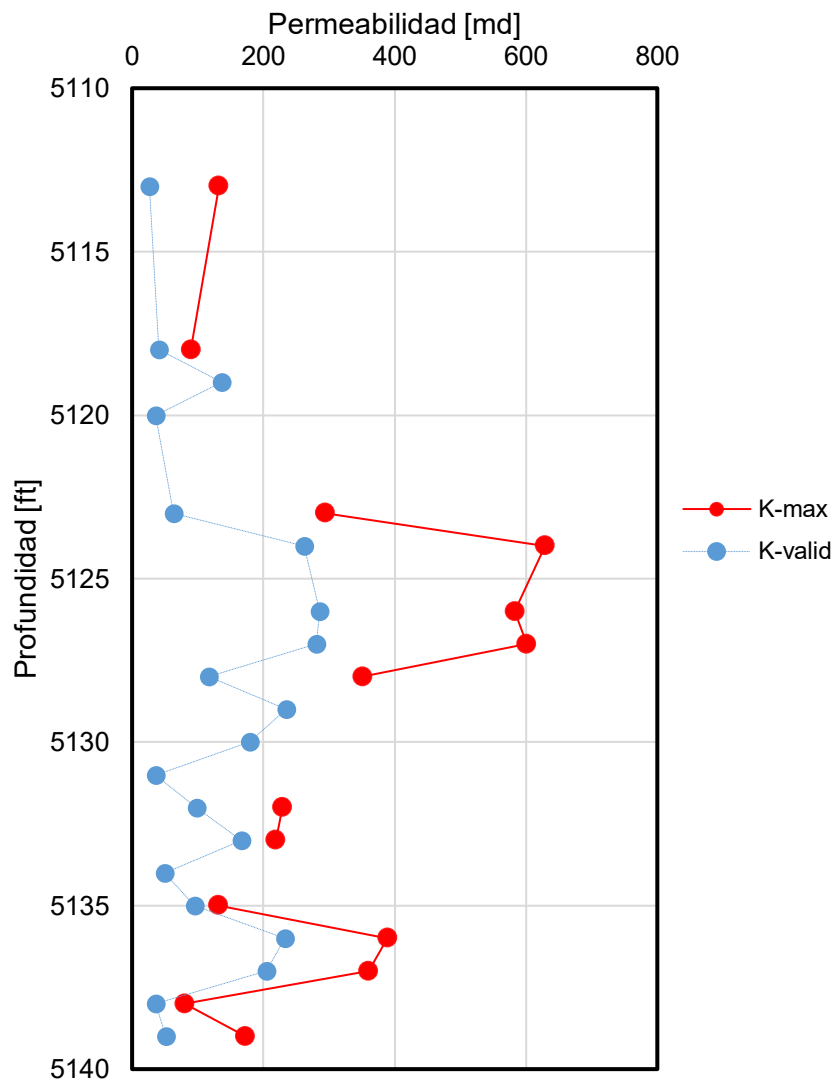


Figura 4.66. Predicción de Permeabilidad del Modelo 1 en el Pozo de Prueba (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).

Modelo 2 (Tangente Hiperbólica – Sigmoide)

En la segunda red, con 8 ocho nodos en la capa de entrada, entre 10 y 270 nodos en la capa oculta y un nodo en la capa de salida, se usa tangente hiperbólica en capa oculta y sigmoide en la de salida como funciones de activación. El modelo se entrenó a una tasa de aprendizaje de 0.5 con 95000 iteraciones.

En la **Figura 4.67** se presentan los valores del coeficiente de correlación R . La red con 50 nodos (resaltado en rojo) es la que mejores resultados arroja.

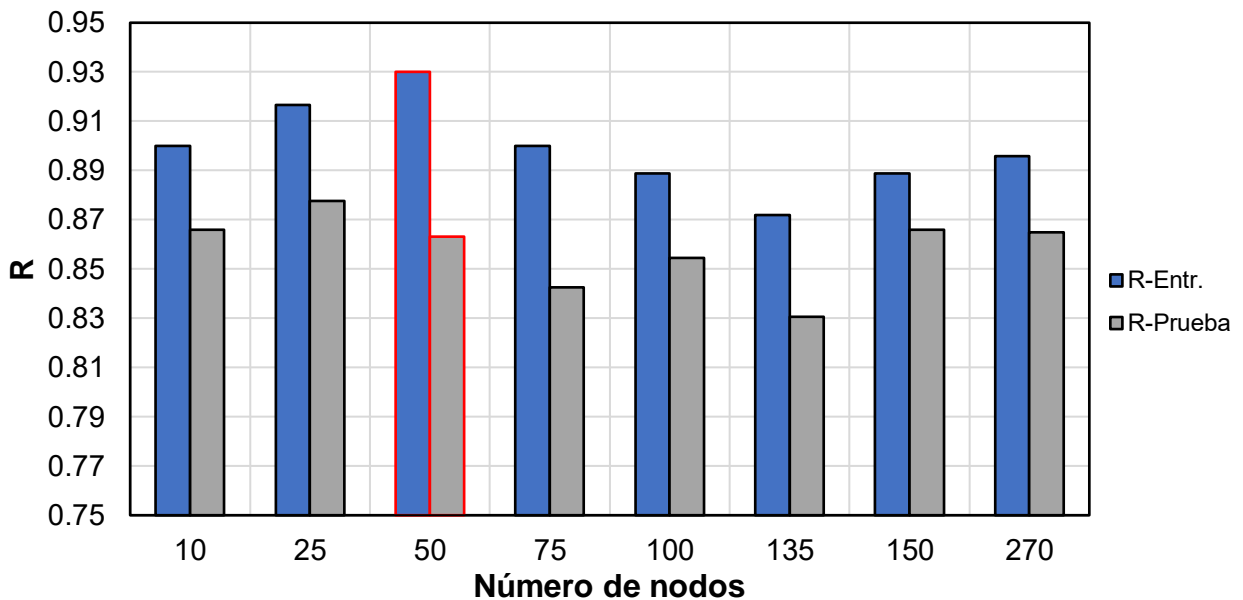


Figura 4.67. Calificación de nodos para el modelo 2 (R-Entr: Coeficiente de correlación del entrenamiento, R-Prueba: Coeficiente de correlación de la prueba).

Resumen del modelo en la **Figura 4.68**. Los errores obtenidos en ambas etapas son visiblemente menores que el modelo anterior.

Entrenamiento	Error de suma de cuadrados	.223
	Error relativo	.095
	Regla de parada utilizada	75000 paso(s) consecutivo(s) sin disminución del error ^a
	Tiempo de entrenamiento	0:01:27.08
Pruebas	Error de suma de cuadrados	.158
	Error relativo	.358
Reserva	Error relativo	.362

Figura 4.68. Resumen del Modelo de Red 1.

A continuación, la **Figura 4.69** muestra la importancia que tienen los atributos de entrada, este segundo modelo la saturación de agua es la variable de mayor importancia seguida de la profundidad y la porosidad.

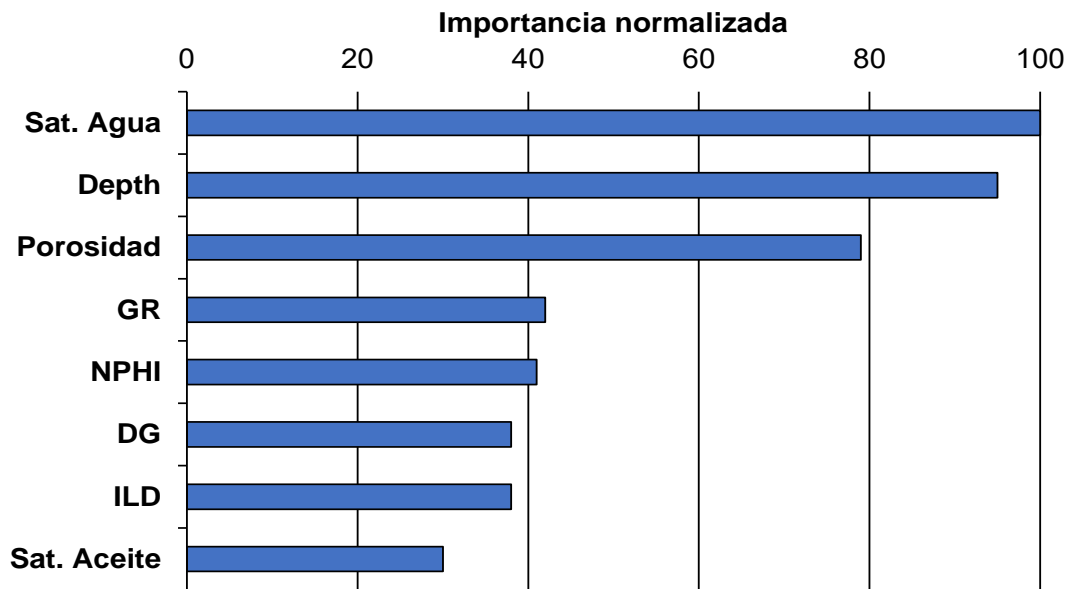


Figura 4.69. Importancia de los Atributos de Entrada para el Modelo Neuronal 2.

La **Figura 4.70** presenta los valores medidos contra los estimados. El modelo subestima de los valores más altos de permeabilidad, sin embargo, se aprecia una reducción en la dispersión para valores de permeabilidad menores a 250 md.

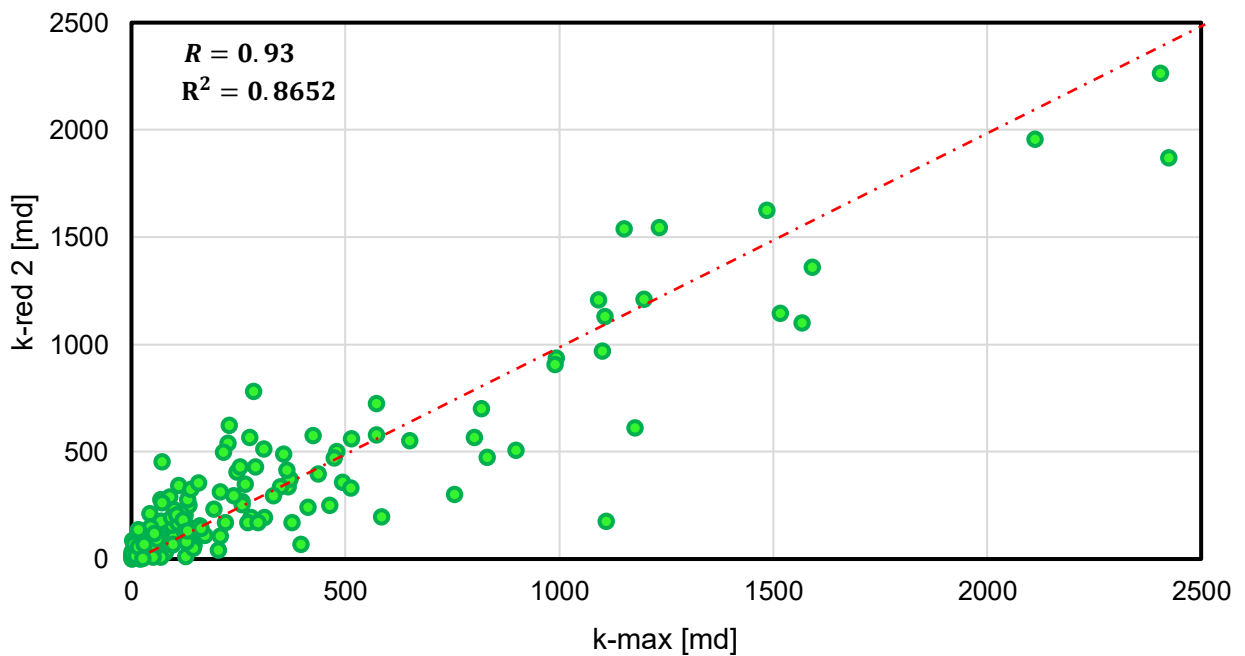


Figura 4.70. Permeabilidad Medida Contra Evaluada para la Red 2 (8 entradas, 50 nodos, 1 salida)

La **Figura 4.71** muestra el poder predictivo de este segundo modelo al comparar los valores aleatoriamente escogidos de la base de datos que no formaron parte del proceso de entrenamiento. El valor de la propiedad más alto se sobreestima aún más que el modelo anterior y existe un menor ajuste en los demás patrones.

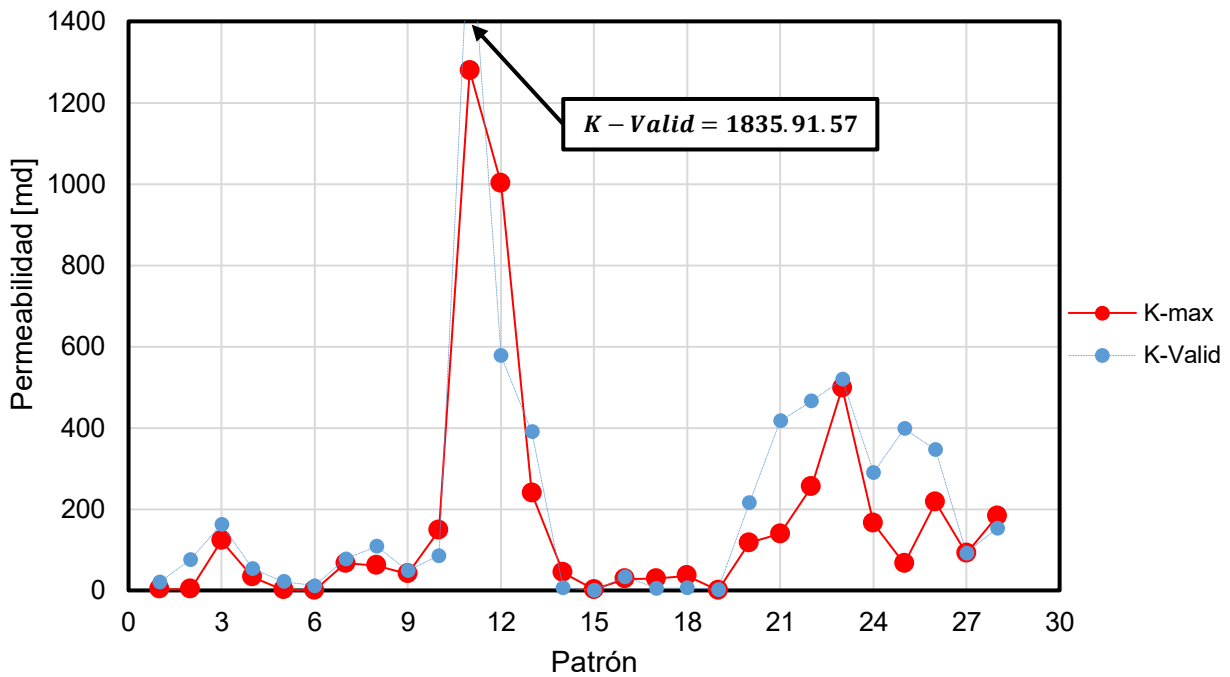


Figura 4.71. Predicción de Permeabilidad por la Red 2. (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).

Para el pozo donde hay valores desconocidos los resultados marcan una tendencia que parece mejorar al anterior modelo (**Figura 4.72**).

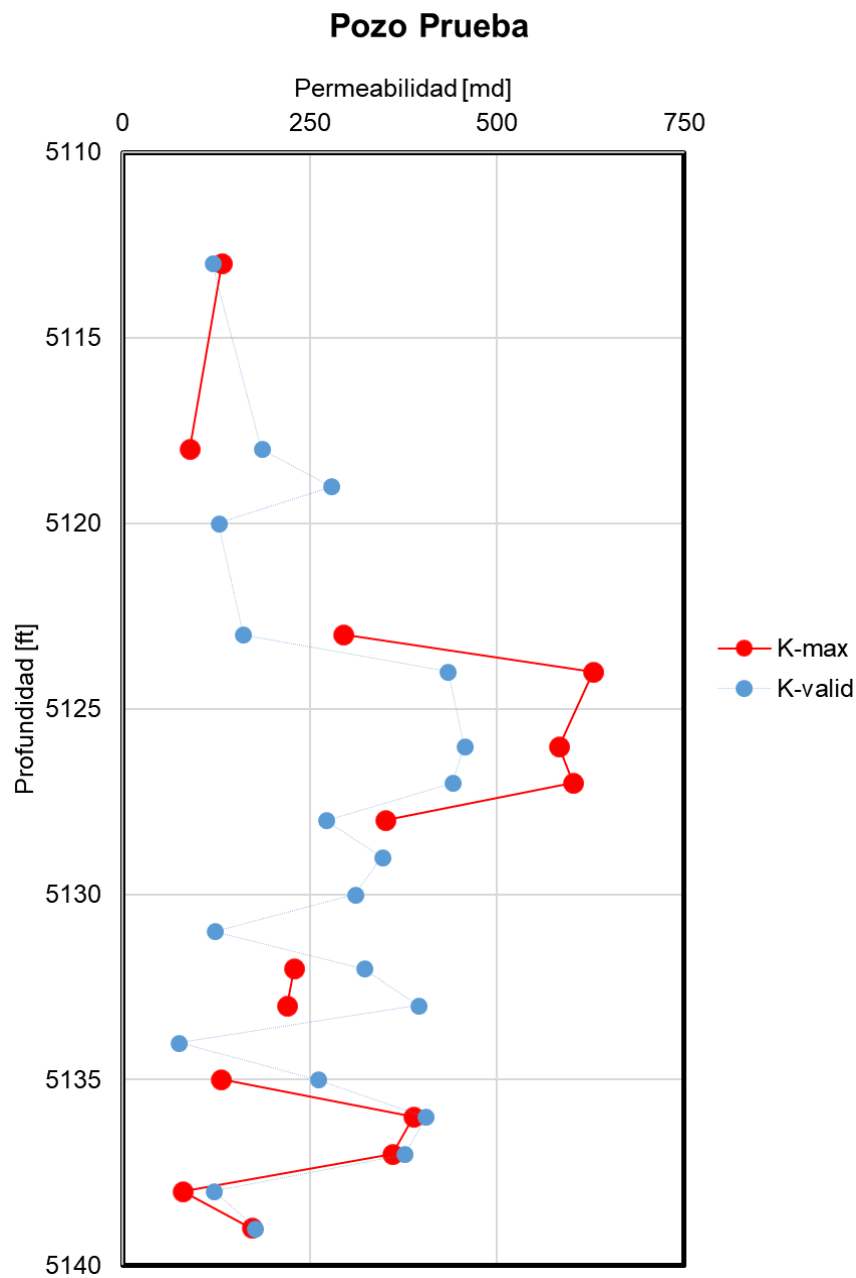


Figura 4.72. Predicción de Permeabilidad del Modelo 2 en el Pozo de Prueba. (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).

Red 3 (sigmoide-sigmoide)

El tercer modelo neuronal, con la misma estructura en capas y número de nodos de salida y entrada usa la función sigmoide como la de activación de la capa oculta y de la capa de salida. El modelo se entrenó con una tasa de aprendizaje de 0.575 con 1,000,000 iteraciones. En la **Figura 4.73** se muestra el gráfico de calificación de nodos. Es evidente

que el caso con los mejores valores del coeficiente de correlación es el de 100 nodos, resaltados en rojo.

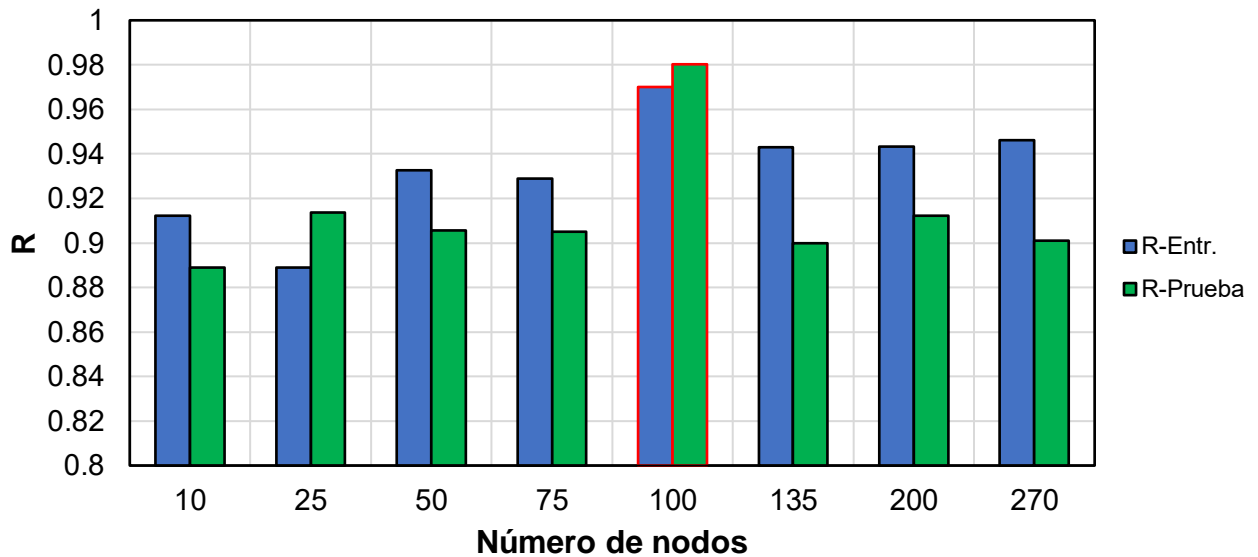


Figura 4.73. Calificación de nodos para el modelo 3. (R-Entr: Coeficiente de correlación del entrenamiento, R-Prueba: Coeficiente de correlación de la prueba).

En la **Figura 4.74** se presenta el resumen paramétrico de la etapa de entrenamiento y prueba de este modelo. Este modelo presenta los errores más pequeños de los tres presentados.

Entrenamiento	Error de suma de cuadrados	.209
	Error relativo	.089
	Regla de parada utilizada	Se ha superado el número máximo de épocas (1000000)
	Tiempo de entrenamiento	0:24:58.68
Pruebas	Error de suma de cuadrados	.205
	Error relativo	.467
Reserva	Error relativo	.294

Figura 4.74. Resumen del Modelo de Red 3.

En la **Figura 4.75** se observa la importancia que tienen los atributos de entrada para este tercer modelo. La variable de mayor importancia es la porosidad, coincidente con lo obtenido en el árbol de regresión. La saturación de agua y la profundidad entre las variables más importantes, así como la densidad de grano. El registro ILD es la variable

menos importante para este modelo. Este modelo es el más adecuado para el cálculo de la permeabilidad.

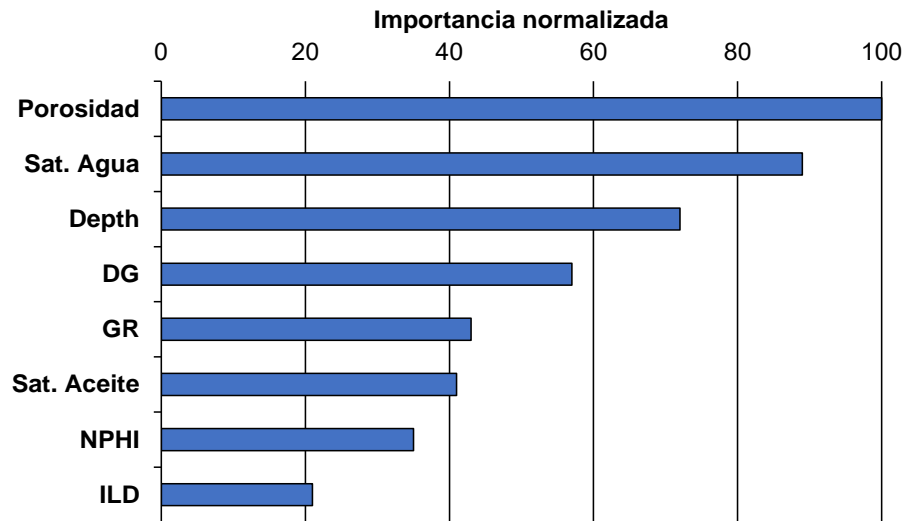


Figura 4.75. Importancia de los Atributos de Entrada para el Modelo Neuronal 3.

En la **Figura 4.76** se presentan las permeabilidades medidas contra las evaluadas y los valores resultantes de los coeficientes de correlación R y R^2 . Se aprecia que los puntos se encuentran más concentrados sobre la línea roja punteada, es decir, la diferencia entre el valor arrojado por la red y el valor real de permeabilidad son más cercanos.

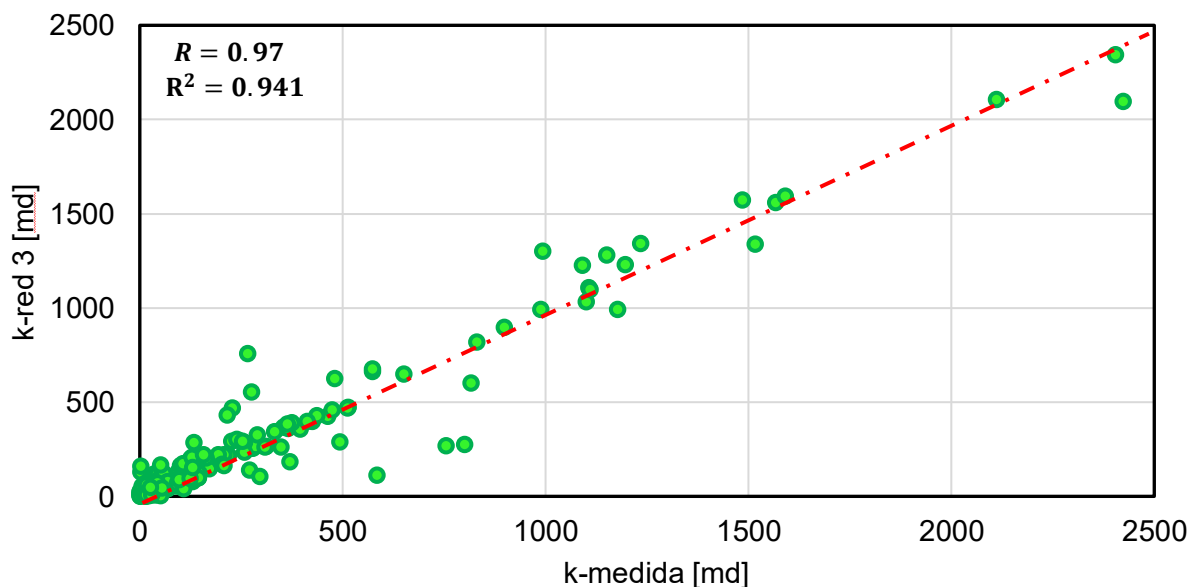


Figura 4.76. Permeabilidad Medida Contra Evaluada para la Red 3.

En la **Figura 4.77** se muestra la capacidad de predicción de este tercer modelo aplicado a los valores aleatoriamente escogidos como validación. Este tercer modelo es mejor en capacidad predictiva incluso para acertar en el pico más grande de permeabilidad mostrado en el patrón 11 (1264 md) y en el valor más pequeño del patrón 2 (0.4 md).

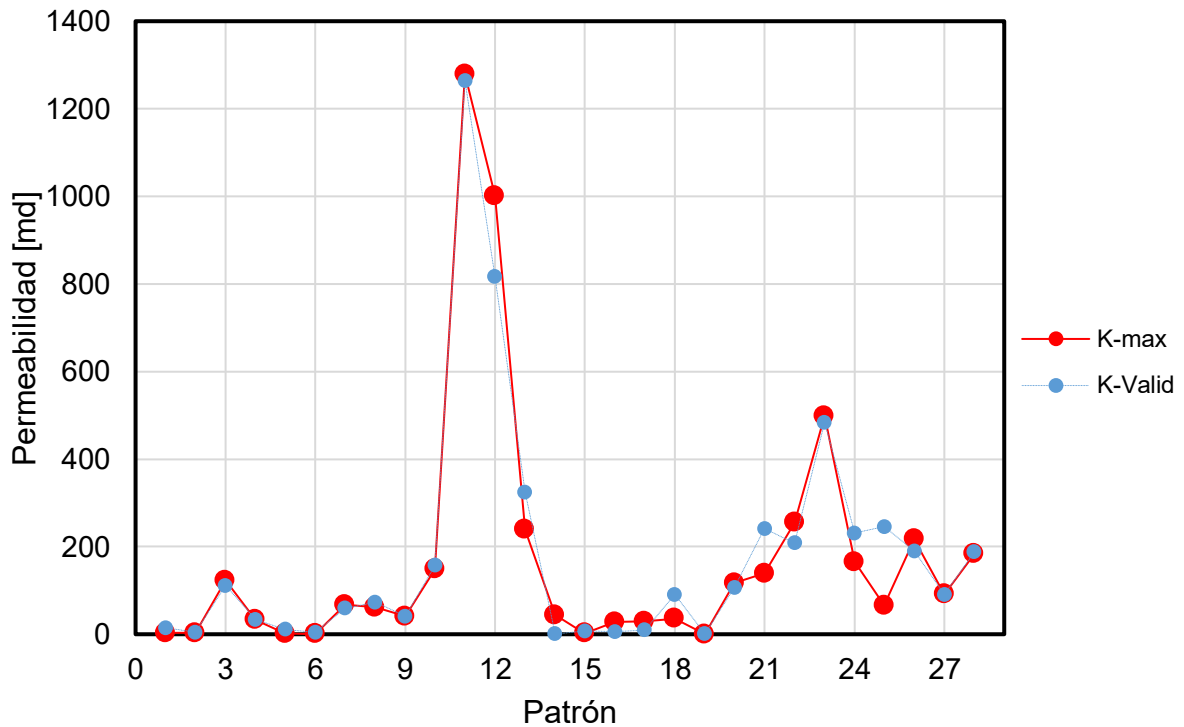


Figura 4.77. Predicción de Permeabilidad por la Red 3. (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).

Para el pozo de prueba (**Figura 4.78**) este modelo presenta mejores predicciones para las secciones con medición y una tendencia más clara y natural en las que no se tiene.

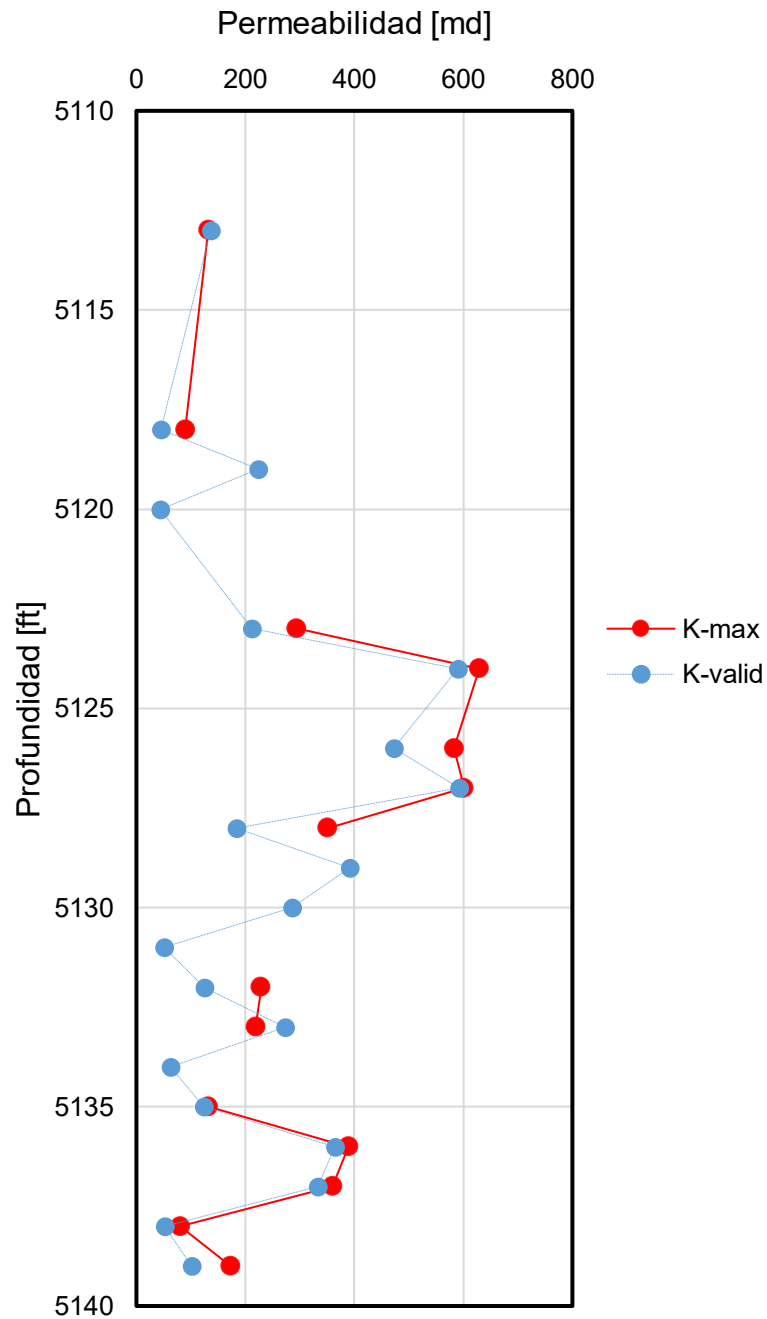


Figura 4.78. Predicción de Permeabilidad del Modelo 3 en el Pozo de Prueba. (k-max: permeabilidad medida, k-valid: permeabilidad aproximada por el modelo).

Se presentan a continuación las comparaciones entre valores de permeabilidad real contra los obtenidos con el modelo neuronal 3 y al árbol de regresión (Figuras 4.79 a 4.85).

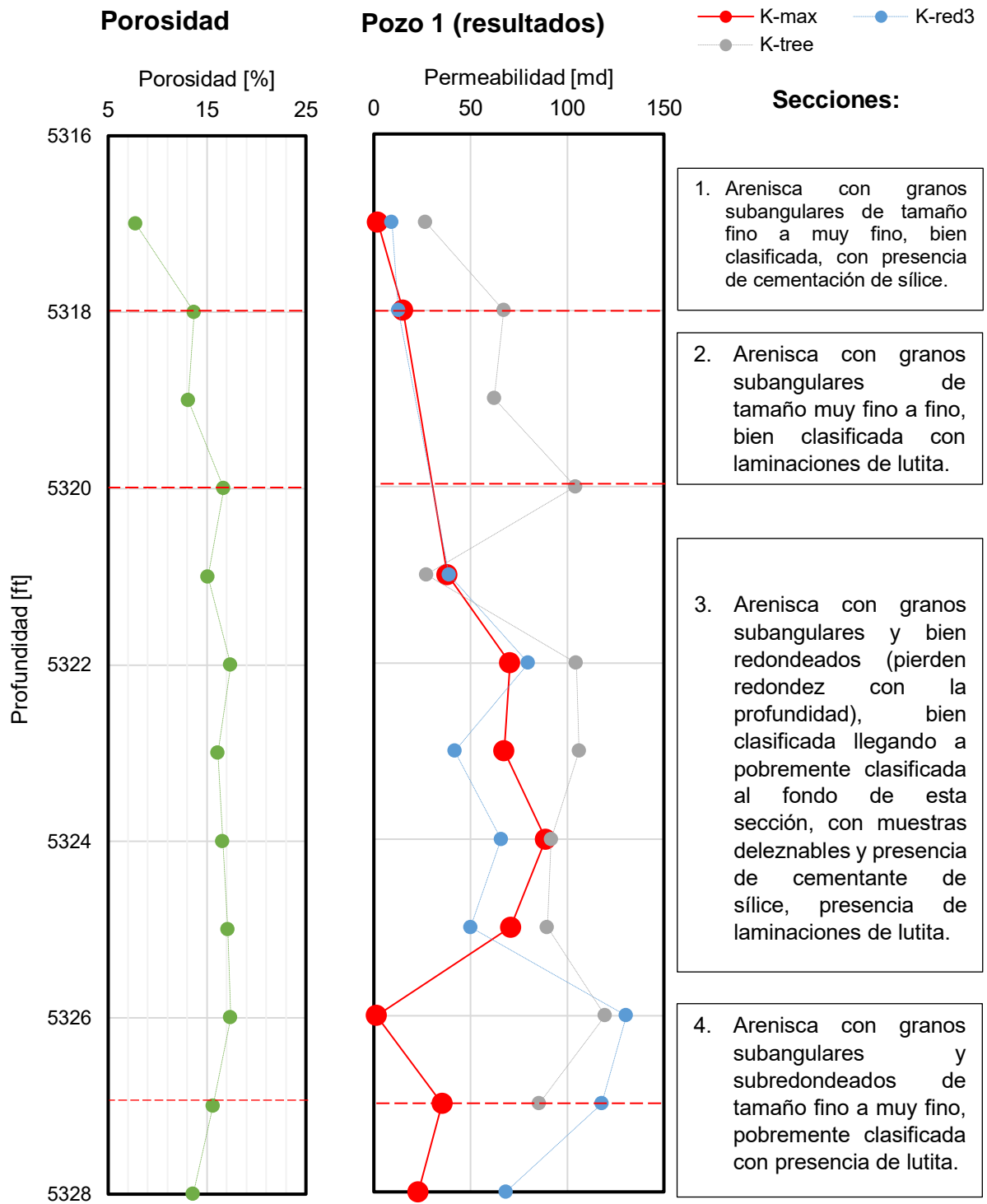


Figura 4.79. Permeabilidad Real y Estimada por la Red 3 en el Pozo 1.

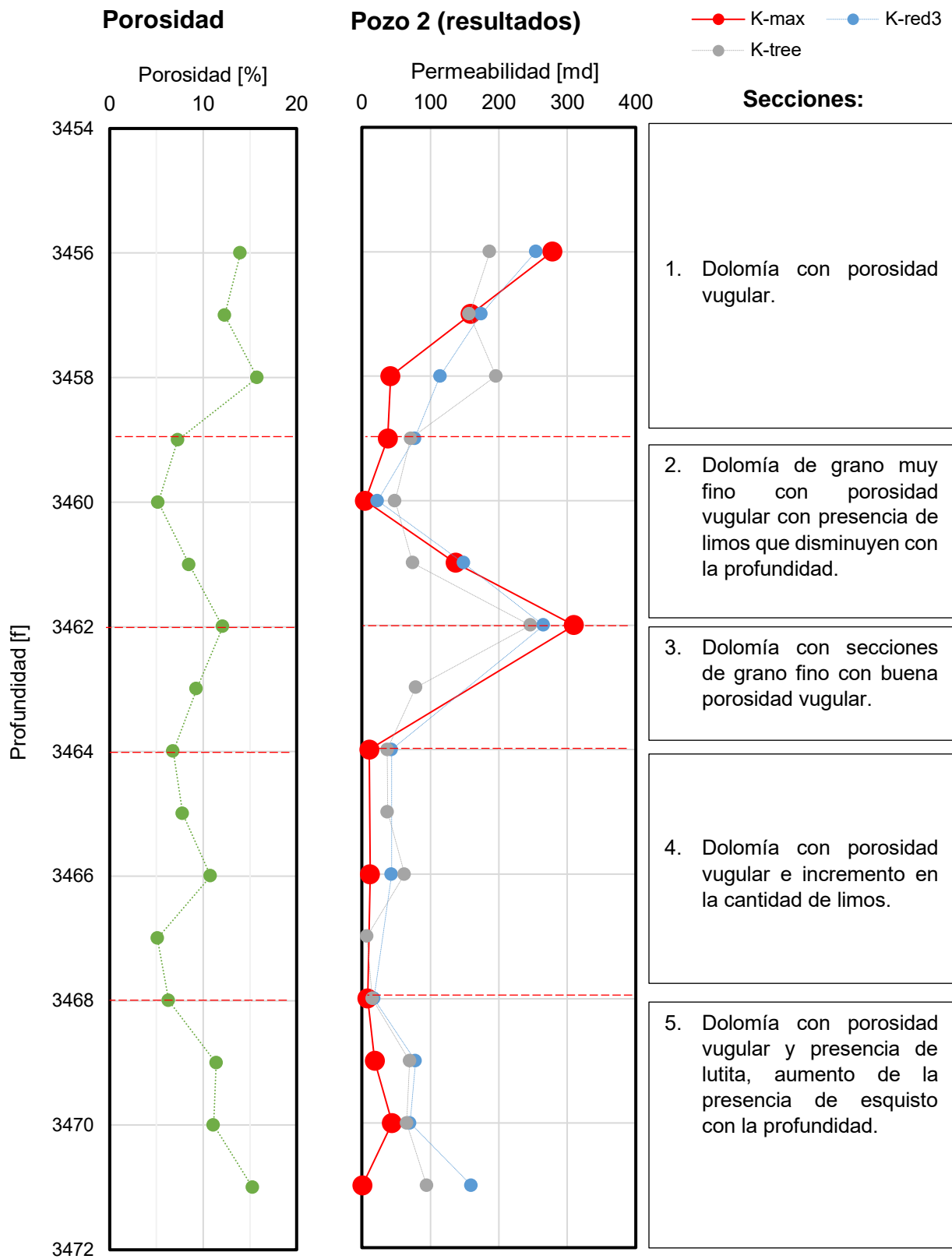


Figura 4.80. Permeabilidad Real y Estimada por la Red 3 en el Pozo 2.

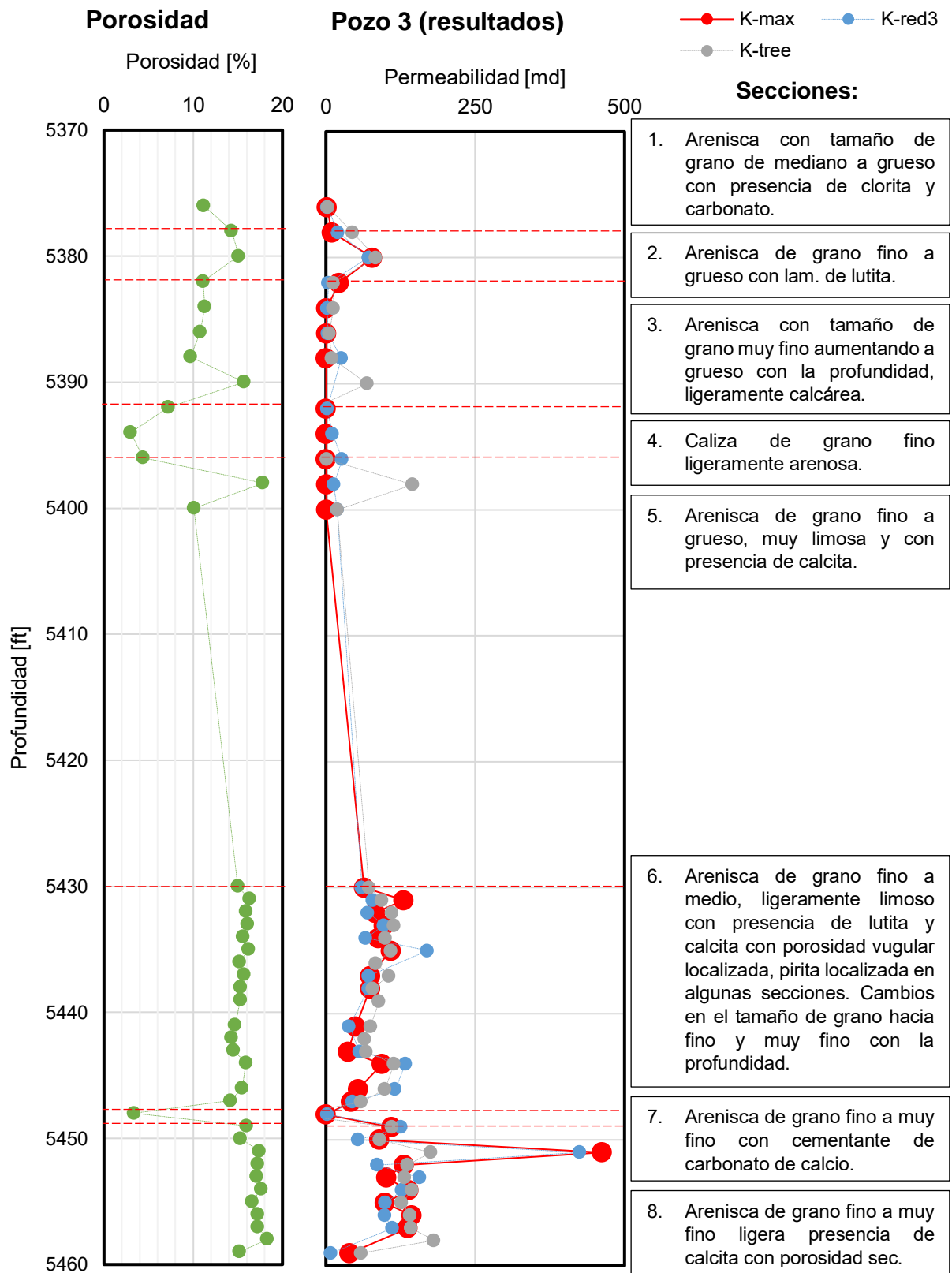


Figura 4.81. Permeabilidad Real y Estimada por la Red 3 en el Pozo 3.

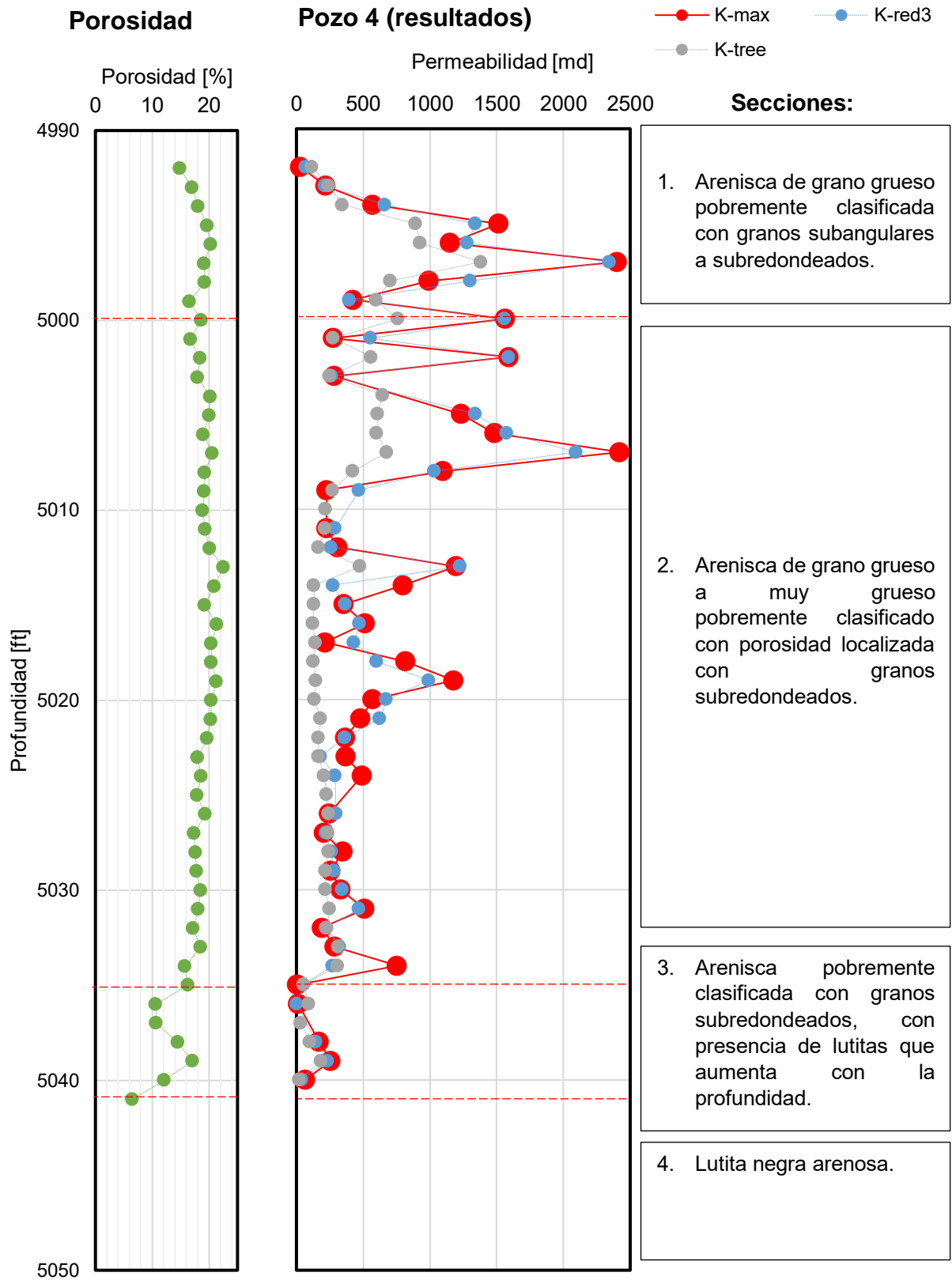


Figura 4.82. Permeabilidad Real y Estimada por la Red 4 en el Pozo 4.

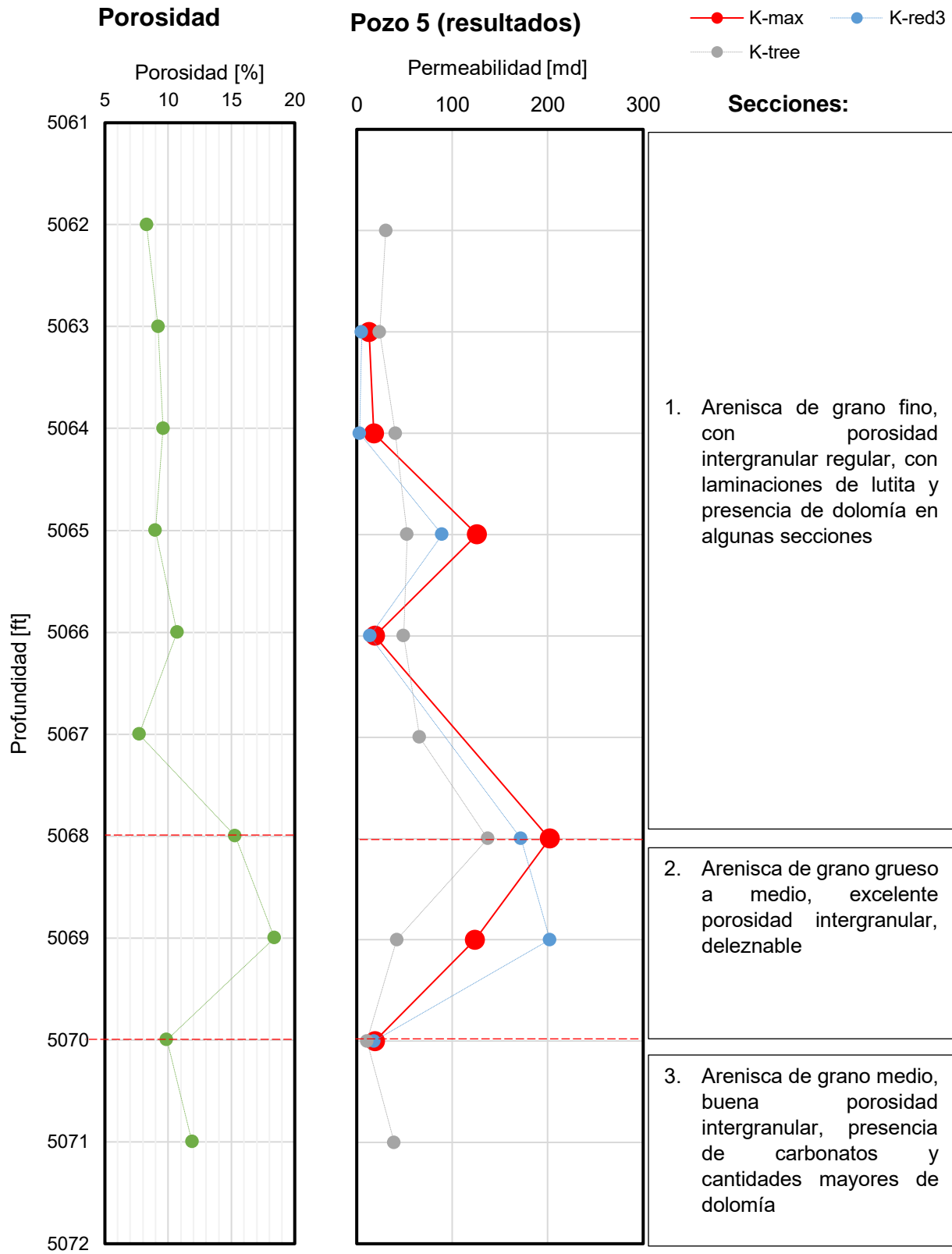


Figura 4.83. Permeabilidad Real y Estimada por la Red 3 en el Pozo 5.

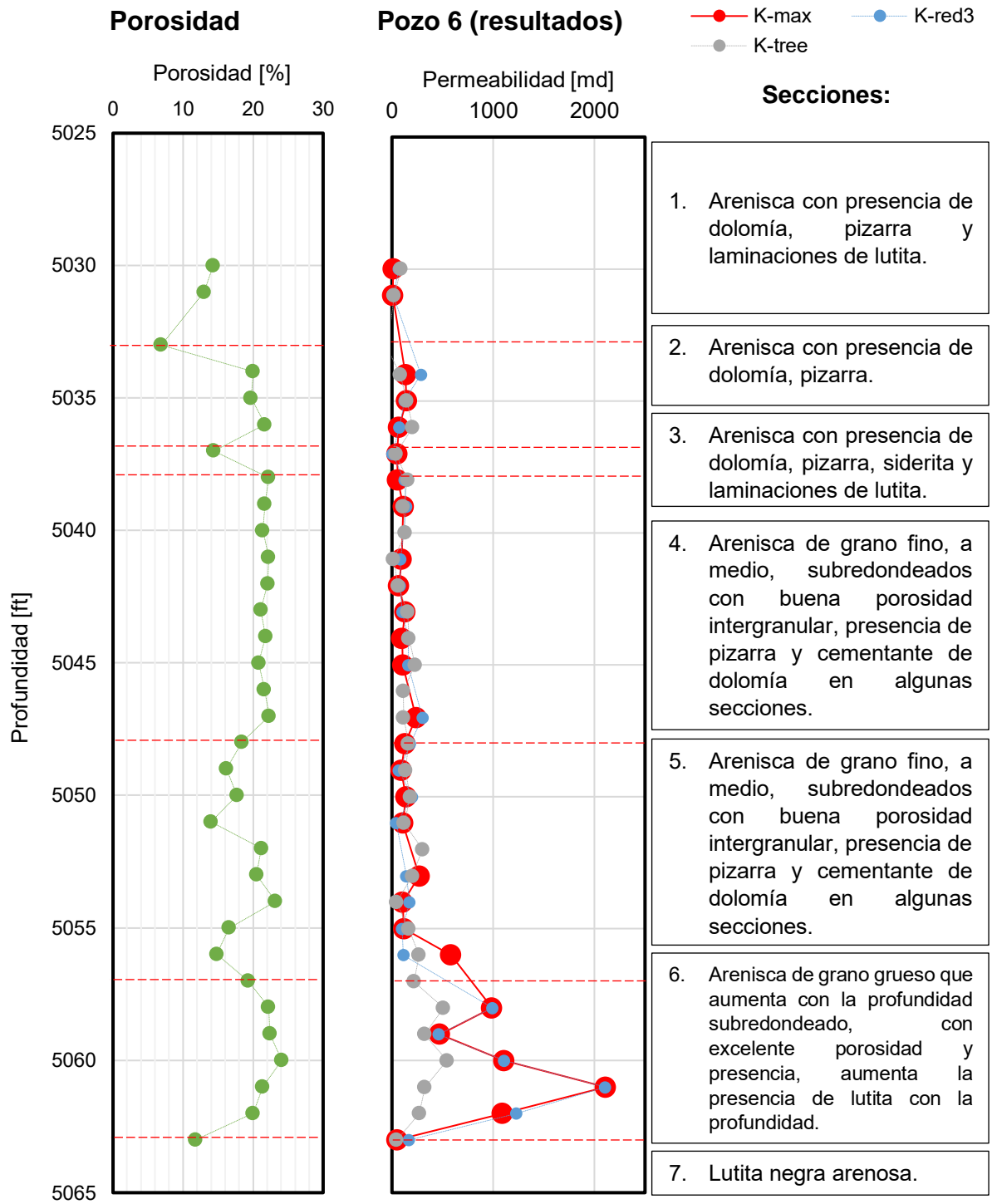


Figura 4.84. Permeabilidad Real y Estimada por la Red 3 en el Pozo 6.

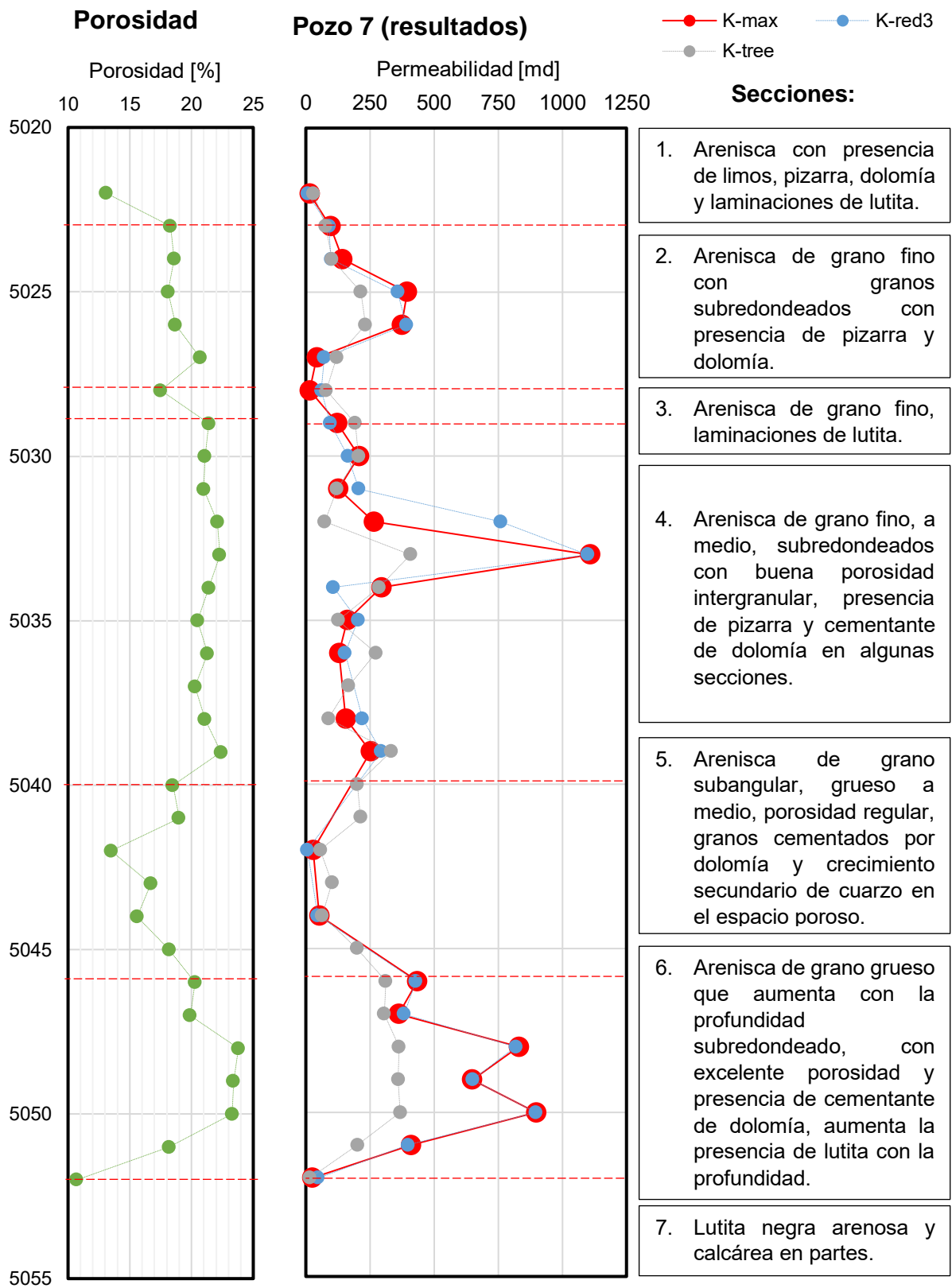


Figura 4.85. Permeabilidad Real y Estimada por la Red 3 en el Pozo 7.

Finalmente, la topología ganadora se muestra en la **Figura 4.86**.

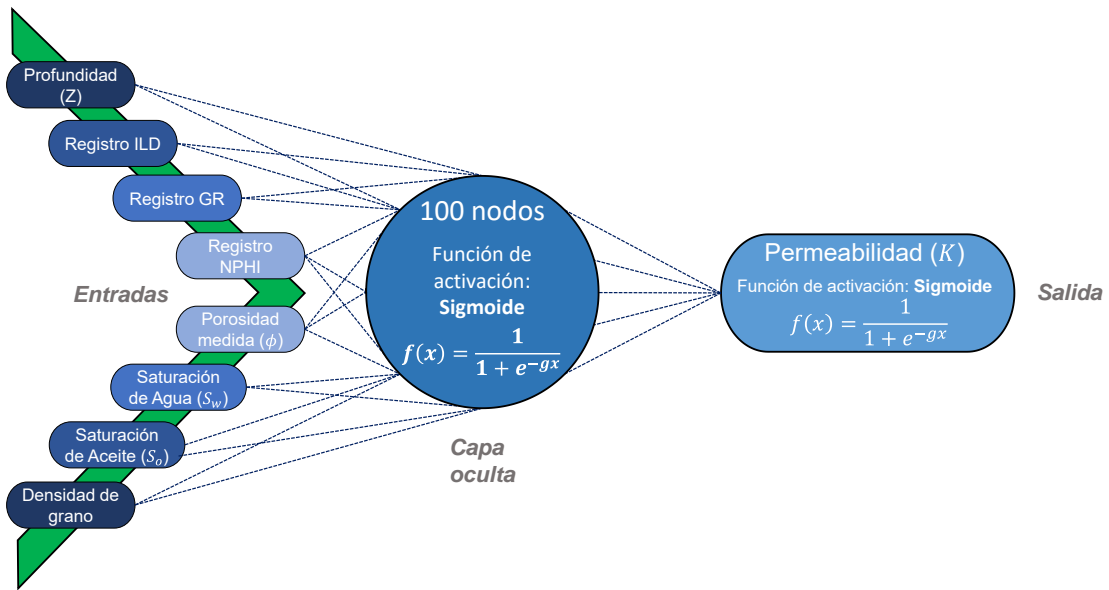


Figura 4.86. Topología del Modelo de Red 3.

Los resultados de la estimación de permeabilidad para dos pozos y su explicación con los factores obtenidos en CD se muestran en las **Figuras 4.87 y 4.88**.

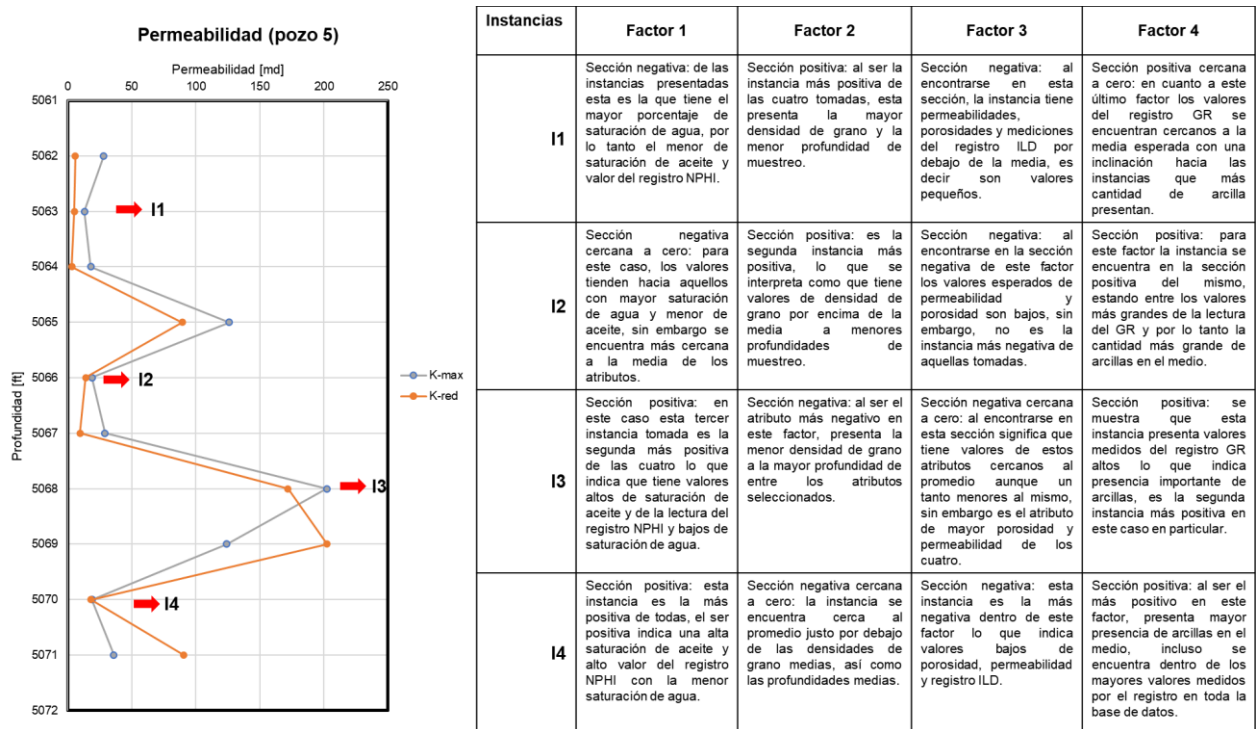


Figura 4.87. Análisis de Resultados de Redes y Análisis Factorial para el Pozo 5.

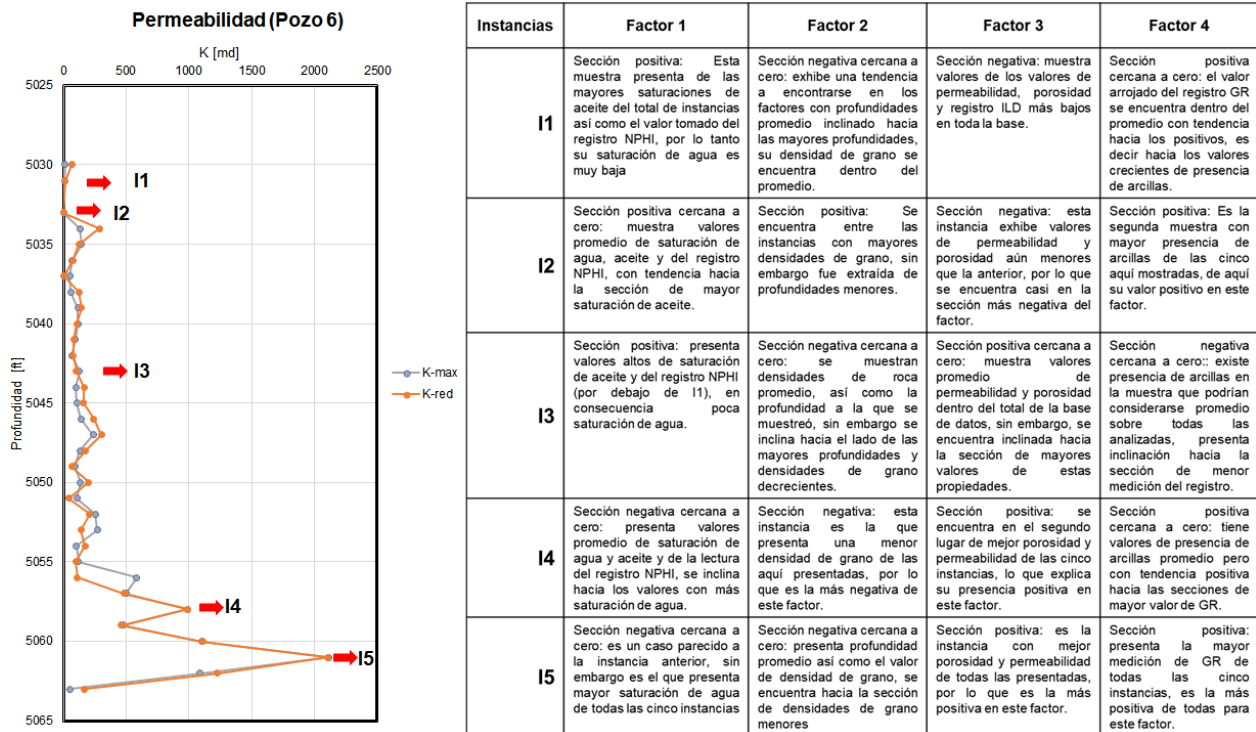


Figura 4.88. Análisis de Resultados de Redes y Análisis Factorial para el Pozo 6.

De este análisis se reconoce que:

- los núcleos con mayores saturaciones de agua presentan los errores más grandes en la predicción del modelo neuronal, mientras que aquellos con saturaciones de aceite mayores o cercanas a la mediana generan mejores ajustes entre lo medido y lo calculado,
- las instancias con mediciones altas del registro GR parecen ser poco sensibles a los valores de las otras entradas, por lo que la permeabilidad se explica con los cambios debido a la diagénesis únicamente.

4.4.4 Resultados: Modelos Semi-empíricos y comparación

Los resultados de la red más exitosa se comparan con valores obtenidos con modelos semi-empíricos. En la **Tabla 4.6** se enlistan los modelos semi-empíricos utilizados, así como la ecuación asociada a éstos, además de sus características y limitaciones de uso. Cabe aclarar que, del extenso número de modelos existentes, se utilizaron aquellos que

podieran ser aplicables según las particularidades geológicas de la zona de estudio y las variables disponibles para el cálculo de la permeabilidad. A continuación, se presentan los perfiles definidos con CART y Redes Neuronales y los valores asociados a cada modelo convencional (**Figuras 4.89 a 4.93**).

Tabla 4.6. Modelos Semi-empíricos Utilizados.

Modelo	Año	Ecuación	Características/Limitaciones
Timur	1968	$K = 0.136 \frac{\phi^{4.4}}{S_{wi}^2}$	<ul style="list-style-type: none"> • No incluye medición alguna del tamaño de garganta de poro. • Funciona de manera óptima para formaciones de areniscas relativamente limpias y homogéneas. • Es necesario ajustar el exponente 4.4 de la porosidad para una buena aproximación según las características del medio
Coates	1974	$K^{1/2} = 100 \frac{\phi^2(1-S_{wirr})}{S_{wirr}}$	<ul style="list-style-type: none"> • Debe respetarse el rango de agua irreductible en la formación. • Estimación buena en formaciones relativamente limpias y homogéneas. • La metodología para estimar la saturación irreductible de agua podría fallar en función de la alta heterogeneidad del medio.
Pape	1999	$K = 0.031\phi + 7.463\phi^2 + 0.191(10\phi)^{10}$	<ul style="list-style-type: none"> • Aproximación general para el cálculo de la permeabilidad a escala de núcleo. • La estructura interna del sistema poroso se basa en un modelo fractal. • Puede ajustarse para funcionar en el rango de areniscas limpias hasta con una gran cantidad de lutitas.

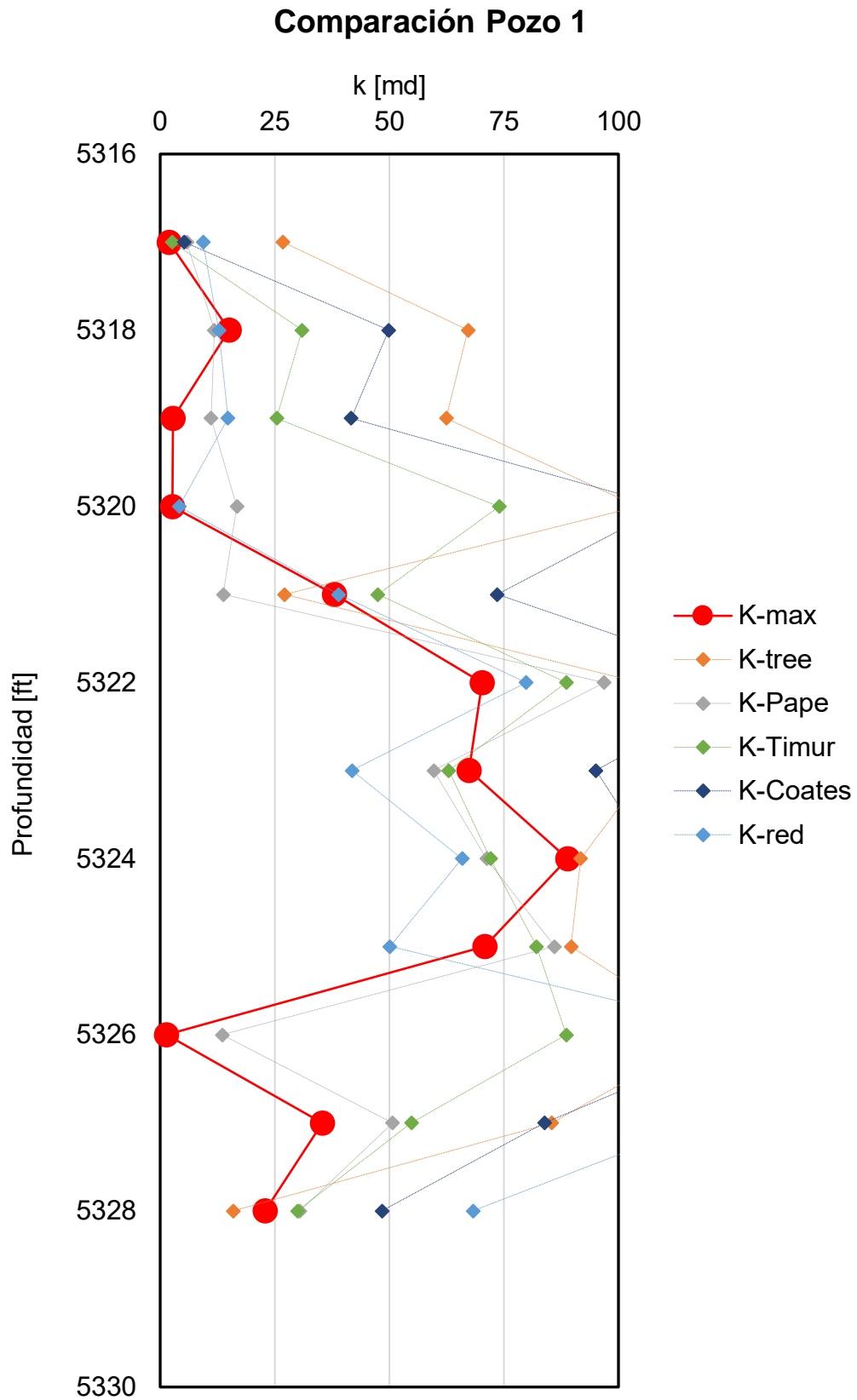
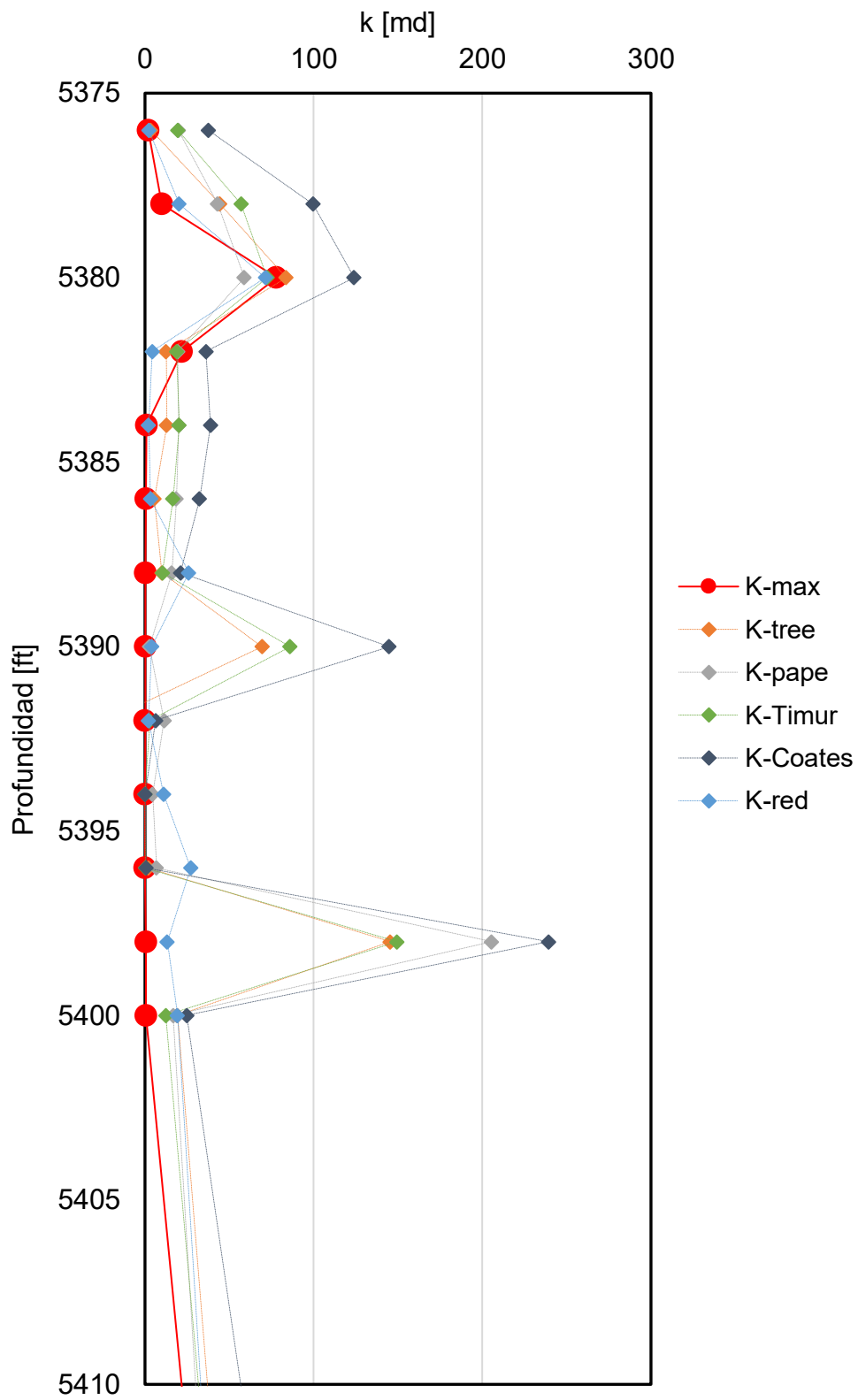


Figura 4.89. Comparación de Resultados para el Pozo 1.

Comparación Pozo 3 (P1)



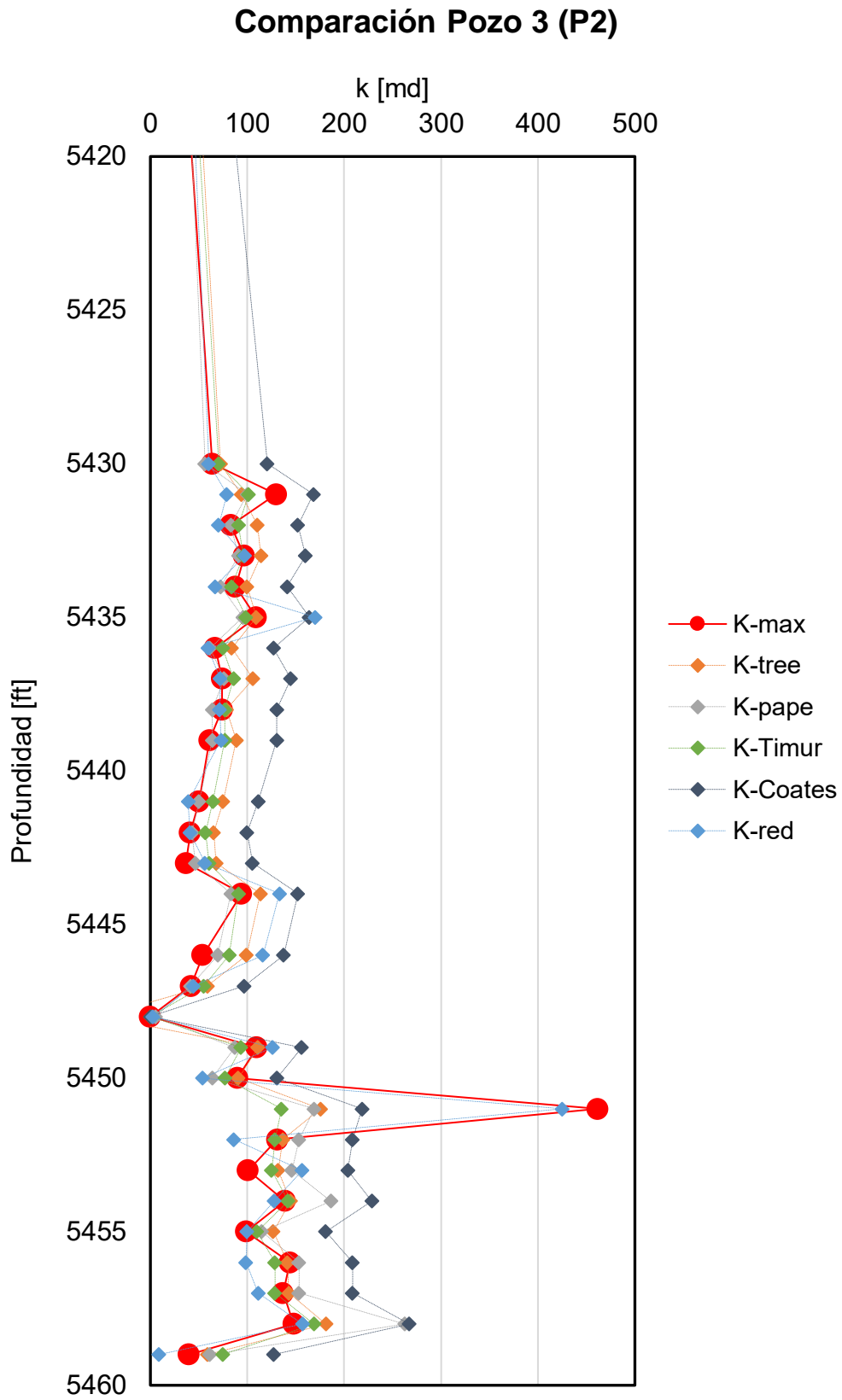


Figura 4.90. Comparación de Resultados para el Pozo 3.

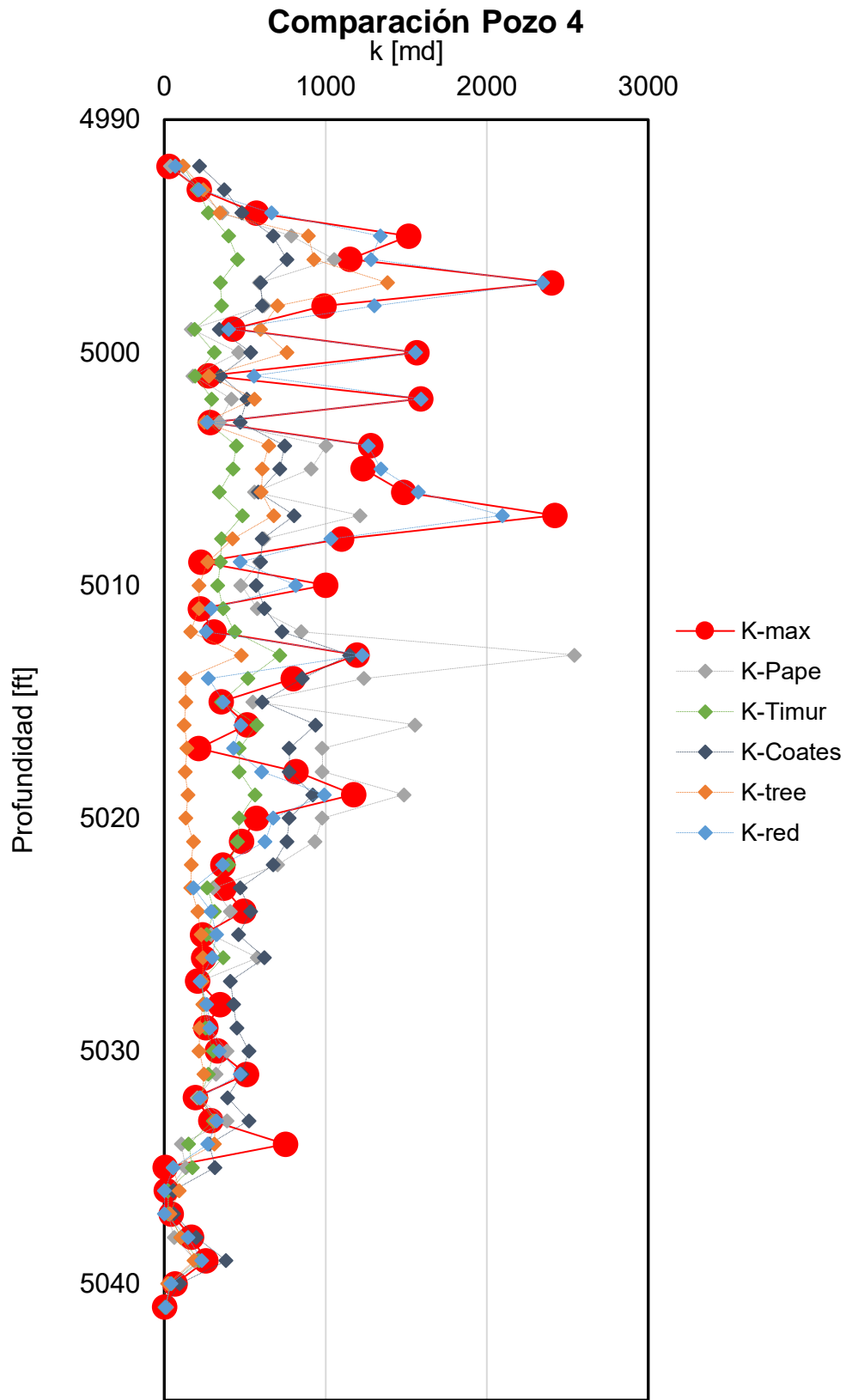


Figura 4.91. Comparación de Resultados para el Pozo 4.

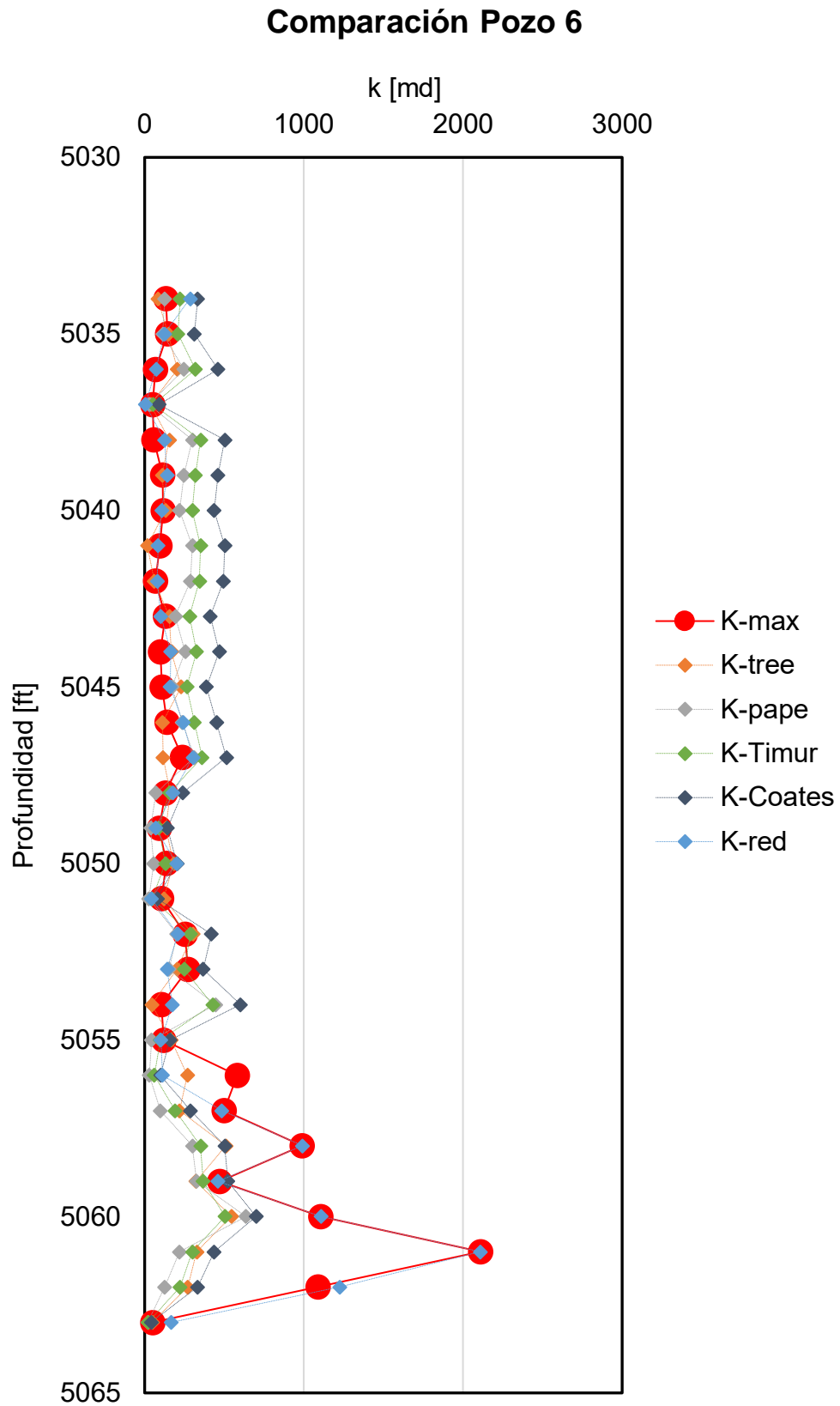


Figura 4.92. Comparación de Resultados para el Pozo 6.

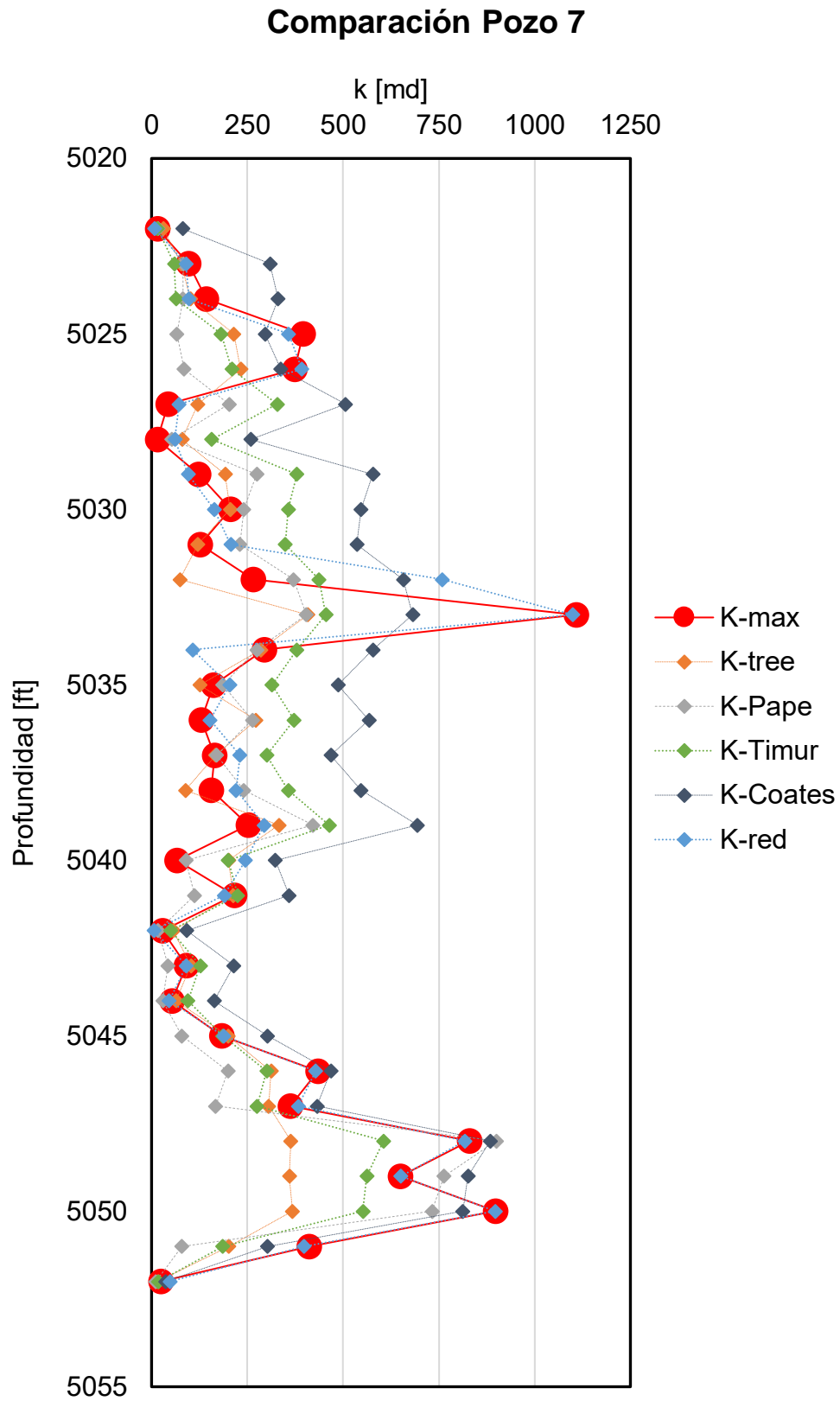


Figura 4.93. Comparación de Resultados para el Pozo 7.

Las **Figuras 4.94 a 4.98** muestran el comportamiento para cada uno de los cinco modelos cuando los valores de permeabilidad estimada se comparan directamente con aquellos medidos de núcleos, donde se obtuvo además el valor de R y R^2 para cada uno de éstos.

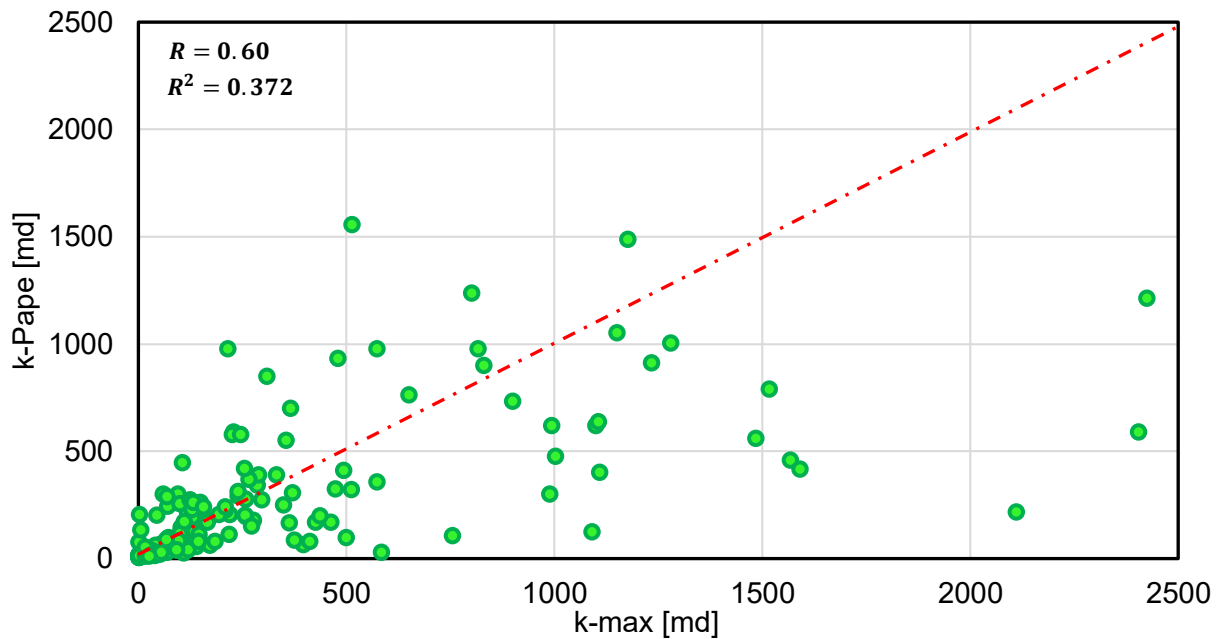


Figura 4.94. Permeabilidad Medida Contra Evaluada para el Modelo de Pape.

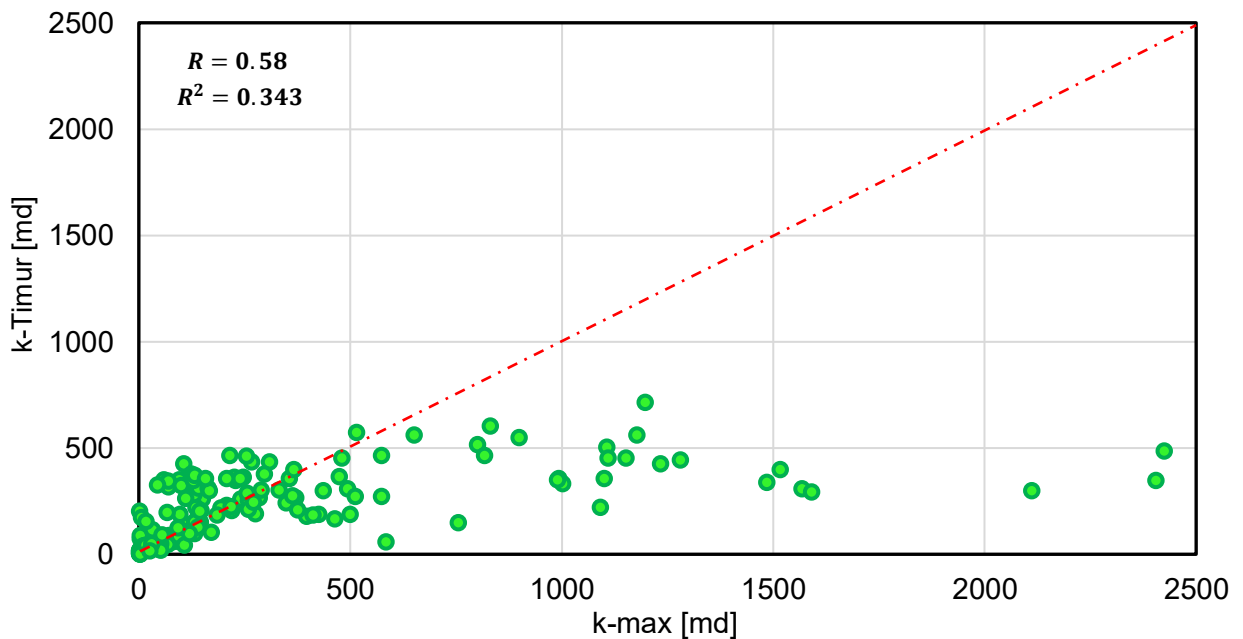


Figura 4.95. Permeabilidad Medida Contra Evaluada para el Modelo de Timur.

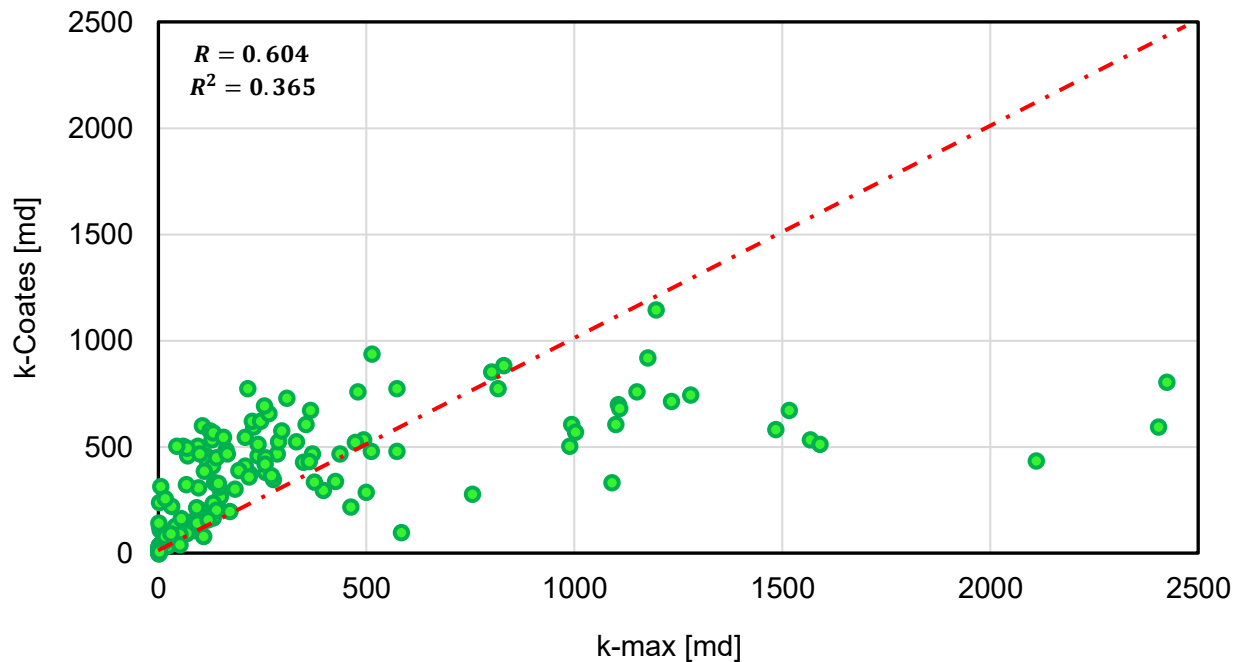


Figura 4.96. Permeabilidad Medida Contra Evaluada para el Modelo de Coates.

Es evidente que los modelos semi-empíricos están lejos de aproximarse al valor medido de permeabilidad para los pozos estudiados. A continuación, la **Tabla 4.7** resume la exactitud en la estimación de la permeabilidad para cada modelo.

Tabla 4.7. Resumen de los Valores de R y R^2 de los Modelos para el Cálculo de la Permeabilidad.

Modelo	R	R^2
Pape (1968)	0.60	0.372
Timur (1981)	0.58	0.343
Coates (1999)	0.60	0.365
CART (M5P)	0.83	0.69
Red 3	0.97	0.94

Finalmente, la **Figura 4.97** muestra los resultados de utilizar estos modelos para predecir los valores de permeabilidad en el pozo de prueba.

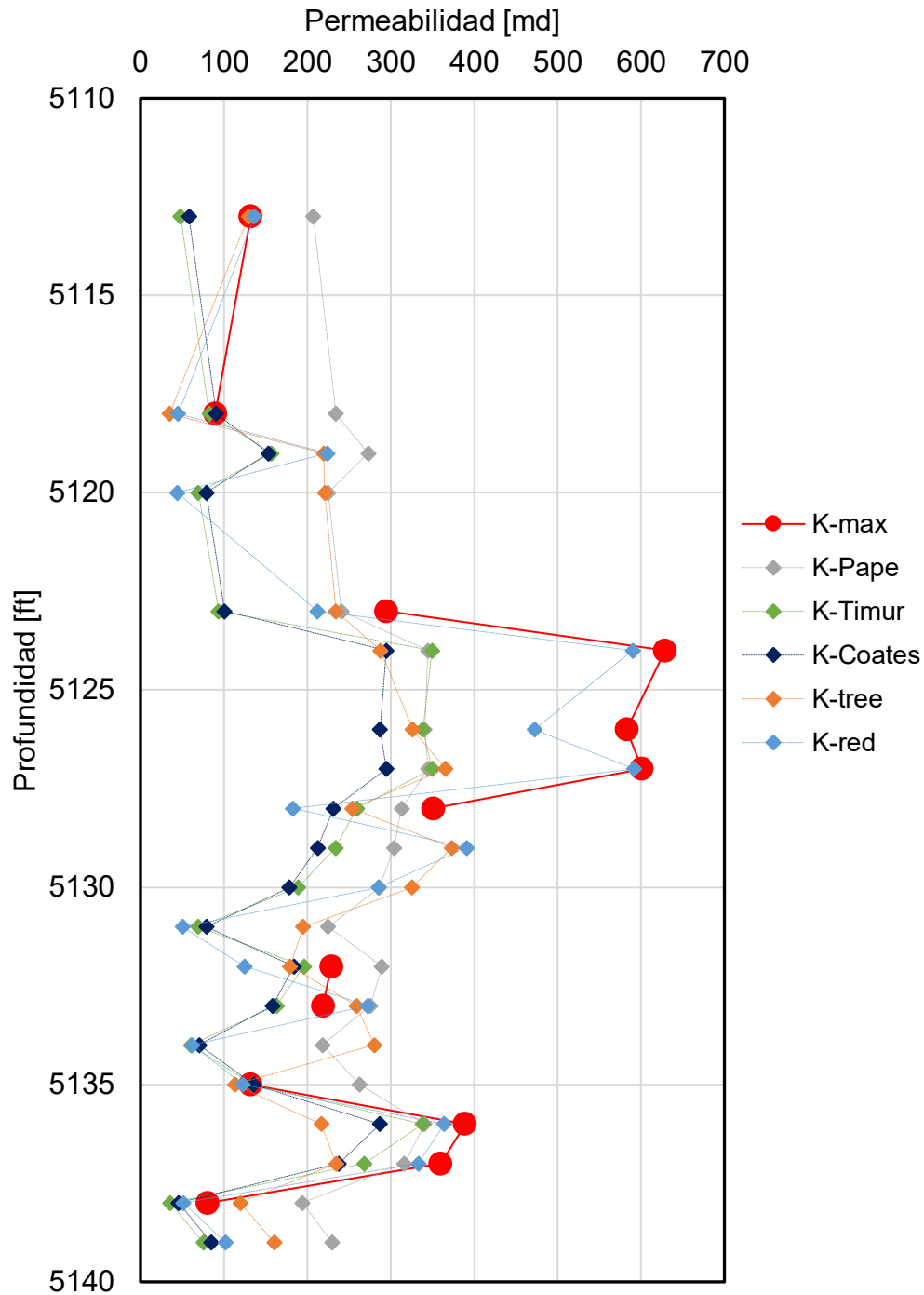


Figura 4.97. Resultado de Permeabilidad en el Pozo Prueba.

Aunque Timur es el mejor modelo para este caso entre los tres modelos convencionales, el árbol de regresión y el modelo neuronal número 3 realizan un trabajo sobresaliente entre éstos.

Capítulo 5: Conclusiones y recomendaciones

Sobre la Ciencia de Datos y las Redes Neuronales para estimar la Permeabilidad

Con base en los resultados obtenidos en esta investigación, se puede concluir que el conocimiento inmerso en las grandes Bases de Datos Relacionales generadas históricamente por la industria petrolera puede ser extraído a través de la aplicación de una atractiva batería de algoritmos informáticos, etiquetados como Ciencia de Datos.

La organización de datos intuitiva dentro de tablas ha migrado hacia esquemas de bases en los que la precisión y la consistencia de las informaciones permiten interpretaciones sólidas y eficientes que respaldan la toma de decisiones, sin embargo, el uso de técnicas de la CD obliga al analista a poseer habilidades en importación y limpieza de datos, análisis exploratorio y aprendizaje máquina.

Los análisis con CD brindaron una solución integral a la necesidad de reconocer comportamientos por propiedad y de todas como un conjunto, permitiendo además ir más allá de predicciones de valor. Los señalamientos de la CD (por ejemplo, en el análisis factorial) permite contar con respaldos no subjetivos para la toma de decisiones en el contexto rápido y heterogéneo de la administración y análisis de datos modernos. Además, el uso de las herramientas de CD trabajadas como un conjunto que se complementa entre si dentro de la metodología presentada permitió llegar a un mejor conocimiento y entendimiento del medio, su heterogeneidad y particularidades que lo definen y como estos factores tienen un impacto en los cambios de la permeabilidad en el medio, además, facilita el entendimiento de como estas pueden ser manejadas o no por los modelos desarrollados.

Por otro lado, desde una perspectiva petrolera, la predicción generada con el modelo neuronal es muy superior a la producida con los métodos semi-empíricos tradicionalmente usados en la industria. Los procedimientos matemáticos “*rígidos*” (del *hard computing*) tienen además desventajas de uso por la demanda de conocimiento específico *a priori* sobre el medio y las conexiones interparamétricas. La RN, entrenada

de forma directa con la información de campo y laboratorio, se convirtió en un sistema que utiliza las propiedades (varias fuentes y varios formatos) que expresan el fenómeno al *aprender* el comportamiento mostrado por los patrones típicos que se le presentaron. Su carácter de caja negra se disolvió en buena medida con los planteamientos de la CD por lo que la red neuronal presentada en esta investigación es un modelo legible, asequible y eficiente, autoexplicativo y fácil de usar para la interpretación paralela de propiedades de entornos naturales complejos.

Al usar una base de datos amplia (en número de atributos e instancias) el conocimiento extraído es sonoro respecto a la teoría conocida pero además se descubrieron algunos aspectos que hacen más eficiente el manejo del conocimiento sobre el medio para aplicarlo en actividades directas de la industria o en el re-diseño de las campañas de exploración y monitoreo.

Sobre la Permeabilidad estimada con RNs y CD

Siendo la industria petrolera la que cuenta con más años comercialmente rentables en la historia de la humanidad podría pensarse que su ingeniería tiene tal grado de madurez que en pocos sentidos podría escapar del *estancamiento*. Sin embargo, esto no es así. Ejemplos son los asombrosos avances (y continuos) en la tecnología, en la práctica y en la comprensión de distintos temas que persiguen aún el objetivo de producir más, más rápido y de mejor manera (menos daño presente y potencial) hidrocarburos. ¿De qué depende el éxito de estas innovaciones? Fundamentalmente del nivel de comprensión sobre el carácter del yacimiento que contiene los hidrocarburos.

La variabilidad y la distribución de las propiedades dentro de los yacimientos, la llamada caracterización de yacimientos pondera a la permeabilidad como la variable que mayores beneficios genera sobre la explotación eficiente de los medios complejos. En esta investigación se propone una manera holística de describir la forma en que un fluido fluye a través del medio poroso haciendo uso de la mayor cantidad de parámetros simples, económicos y que a su vez signifiquen en las complejas dependencias reconocidas en el fenómeno. Se usa información de mediciones en el laboratorio como en el campo para salvar las inconsistencias intrínsecas a las pruebas y monitoreo.

Con la integración de las variables Densidad de Grano, Registro de Rayos Gamma y Porosidad se toma en cuenta la heterogeneidad de la permeabilidad de los yacimientos (y como condición especial las secciones homogéneas) y se apunta hacia una herramienta que permitiría la definición de la distribución de la permeabilidad durante el período de caracterización de yacimientos. Las correlaciones entre parámetros ayudan a la comprensión de la variación de esta propiedad en las direcciones de importancia y a la determinación de estratos geológicos y procesos de formación.

Finalmente, los análisis realizados para determinar las variables que más peso tienen en los modelos de aproximación de permeabilidad (sean árbol de regresión o redes neuronales) mostraron resultados cuya confiabilidad está respaldada con la teoría ya conocida sobre este fenómeno ya que aquellas variables, como la porosidad, tuvieron un mayor impacto en los modelos, además dichos análisis fueron de utilidad, para reconocer la existencia de posibles sesgos ya sean de origen natural, derivados de las actividades de medición o bien propio de la base de datos, lo que en conjunto es útil para explicar la validez de cada modelo diseñado y junto con los valores aproximados el porqué de la selección de un modelo sobre otro.

Nomenclatura

Símbolo	Significado
$A_{sección}$	Área de la sección del recipiente
A	Área Transversal al Gasto
a_{jh}	Coefficiente de Correlación parcial entre las Variables X_j y X_h
B	Factor de Volumen
c	Constante que depende de propiedades de la roca y el fluido
CD	Ciencia de Datos
CH	Conductividad Hidráulica
CM	Comunicación
CP	Cómputo
DE	Ambiente de datos
e_1, e_2, \dots, e_p	Factores Únicos o Específicos
F_1, F_2, \dots, F_m	Factores Comunes
H_0	Hipótesis Nula
H_1	Hipótesis no-Nula
h	Altura o Grosor de la Formación
I_{jh}	Espesor del factor h
I_{hk}	Incógnitas
INF	Informática
K/K_a	Permeabilidad Absoluta
K_{ef}	Permeabilidad Efectiva al Fluido
K_{eo}	Permeabilidad Efectiva al Aceite
K_g	Permeabilidad al Gas
K_l	Permeabilidad al Líquido
K_{rf}	Permeabilidad Relativa al Fluido
L	Longitud de la Muestra
MG	Administración

m_o	Movilidad
P_i	Presión Inicial
P_m	Presión Media
P_{wf}	Presión de Fondo Fluyente
P_{ws}	Presión de Fondo Estática
q	Gasto de Fluido
r_{hj}	Coefficiente de Correlación Muestral
R_i/R_{ti}	Resistividad de la Formación a la Saturación de Agua Irreductible
r_{jh}	Coefficiente de Correlación observados entre las Variables X_j y X_h
R_o	Resistividad de la Formación 100% Saturada de agua
R_w	Resistividad de la Salmuera
SL	Sociología
ST	Estadística
S_{wi}	Saturación de Agua Irreductible
TH	Pensamiento
t_p	Tiempo de Producción antes del Cierre
w	Exponente w
\bar{X}	Media o Promedio Aritmético de la Variable
$\frac{\Delta h}{\Delta l}$	Gradiente Hidráulico
ΔP	Caída de Presión
Δt	Tiempo al que se Realizó la Medición de Presión Después del Cierre
μ	Viscosidad del Fluido
μ_f	Viscosidad del Fluido
ρ_{hj}	Coefficiente de Correlación Poblacional
ρ_o	Densidad del Aceite
ρ_w	Densidad de la Salmuera
σ^2	Varianza
ϕ	Porosidad
ϑ_{hj}	Multiplicadores de Lagrange

Referencias

1. Adams, W. (1964). Diagenetic Aspects of Lower Morrowan, Pennsylvanian Sandstones, northwestern Oklahoma. *Am. Assoc. Petroleum Geologists Bull*, 48(9), 1368-1580.
2. Aguilera, A. (2005). *Análisis de tablas de contingencia bidimensionales*. Granada: Universidad de Granada.
3. Alonso, F. (2006). *La porosidad en Rocas Carbonatadas*. Oviedo, España: Universidad de Oviedo.
4. Aminian, K., Ameri, S., Oyerokun, A., & Thomas, B. (2003). Prediction of Flow Units and Permeability Using Artificial Neural Networks. *SPE Western Regional/AAPG* (págs. 1-7). Long Beach: SPE.
5. Andersen, M., & Duncan, B. (2013). Los Núcleos en la Evaluación de Formaciones. *Oilfield Review*, 25(2), 16-27.
6. Andersen, M., & Klemin, D. (2014). Defining and Determining Permeability. *Oilfield Review*, 26(3), 1-2.
7. Anthony, M., & Holden, B. (1998). Cross-validation for binary classification by real-valued functions: theoretical analysis. *In Proc.*, 218-229.
8. API. (1988). *Recommended Practices for Core Analysis* (segunda ed.). Washington: API.
9. Arroyo Carrasco, A. (2007). *Bases Teóricas e Interpretación de Registros Geofísicos de Pozo* (Primera ed.). Ciudad de México, México: Facultad de Ingeniería - Universidad Nacional Autónoma de México.
10. Balan, B., Mohaghegh, S., & Ameri, S. (1995). State-of-the-Art in Permeability Determination from Well Log Data: Part 1-A Comparative Study, Model Development. *SPE Eastern Regional Conference & Exhibition.*, (pág. 10). Morgantown.
11. Becerra, J. (?). *Matemáticas Básicas: Estadística Descriptiva*. Ciudad de México: Universidad Nacional Autónoma de México.
12. Berger, R. (2014). *A Scientific Approach to Writing for Engineers and Scientists*. Wiley.
13. Bierman, H., & Fernández, L. (1998). *Game Theory with Economic Applications*. Addison - Wesley.
14. Bornemann, O. (1983). Stratigraphie und Tektonik des Zechsteins im Salzsotock Gorleben auf Grund von Bohrergebnissen. *Z Dtsch Geol Ges*, 119-134.

15. Botset, H. (1933). *The Measurement of Permeability of Porous Media for Homogeneous Fluids* (Vol. 4). New York: Review of Scientific Instruments.
16. Breiman, L., & et al. (1984). *Classification and Regression Trees*. Wadsworth, California.
17. Breiman, L., Friedman, J., & et al. (1984). *Classification and Regression Trees*. Belmont, California, USA: Wadsworth International Group.
18. Brown, M. (2014). *Data Mining for Dummies*. Londres: Wiley.
19. Bubenick, L., Chilingar, G., Cook, H., & Egbert, R. (1983). *Diagenesis in Sediments and Sedimentary Rocks* (Vol. 2). New York: Elsevier Scientific Publishing Company.
20. Caldwell, C., & Boeken, R. (1985). Wireline log Zones and Core Description of Upper Part of the Middle Ordovician Viola Limestone, McClain and McClain SW Fields, Nemaha County, Kansas. *Kansas Geological Survey, Subsurface Geology*(6), 17-35.
21. Cao, L. (2017). Data Science: a Comprehensive Overview. *ACM Comput. Surv.*, ?
22. Carles, M. (2008). *Nuevos Métodos de Análisis Multivariante*. CMC Editions.
23. Charniak, E., Riesbeck, C., & et al. (1987). *Artificial Intelligence Programing* (2a ed.). Lawrence: Lawrence Erlbaum Associates.
24. Coates, G., & Dumanoir, J. (1974). A New Approach to Improved Log-Derived Permeability. *The Log-Analyst*, 17.
25. Darcy, H. (1856). *Les Fontaines Publiques de la Ville de Dijon*. Dijon.
26. de la Fuente Fernández, S. (2011). *Análisis Factorial*. Universidad Autónoma de Madrid.
27. De'ath, G., & Fabricius, K. (2000). Classification and Regression Trees: A Powerful Yet Simple Techinque for Ecological Data Analysis. *Ecology*, 81 (11).
28. Deconick, Z., & et al. (2006). Classification Trees Models for the Prediction of Blood-brain Barrier Passage of Drugs. *Journal of Chemical Information and Modeling*, 46(3), 1410-1419.
29. Dmitrienko, A., & Tamhane, A. (2009). *Multiple Testing Problems in Pharmaceutical Statistics*. Londres: Chapman and Hall/CRC.
30. Earlougher, R. (1977). *Advances in Well Test Analysis* (Vol. 5to). New York: SPE.
31. Exploration and Production Department API. (1998). *Recommended Practices for Core Analysis* (Segunda ed.). D.C.: American Petroleum Institute Publishing Services.

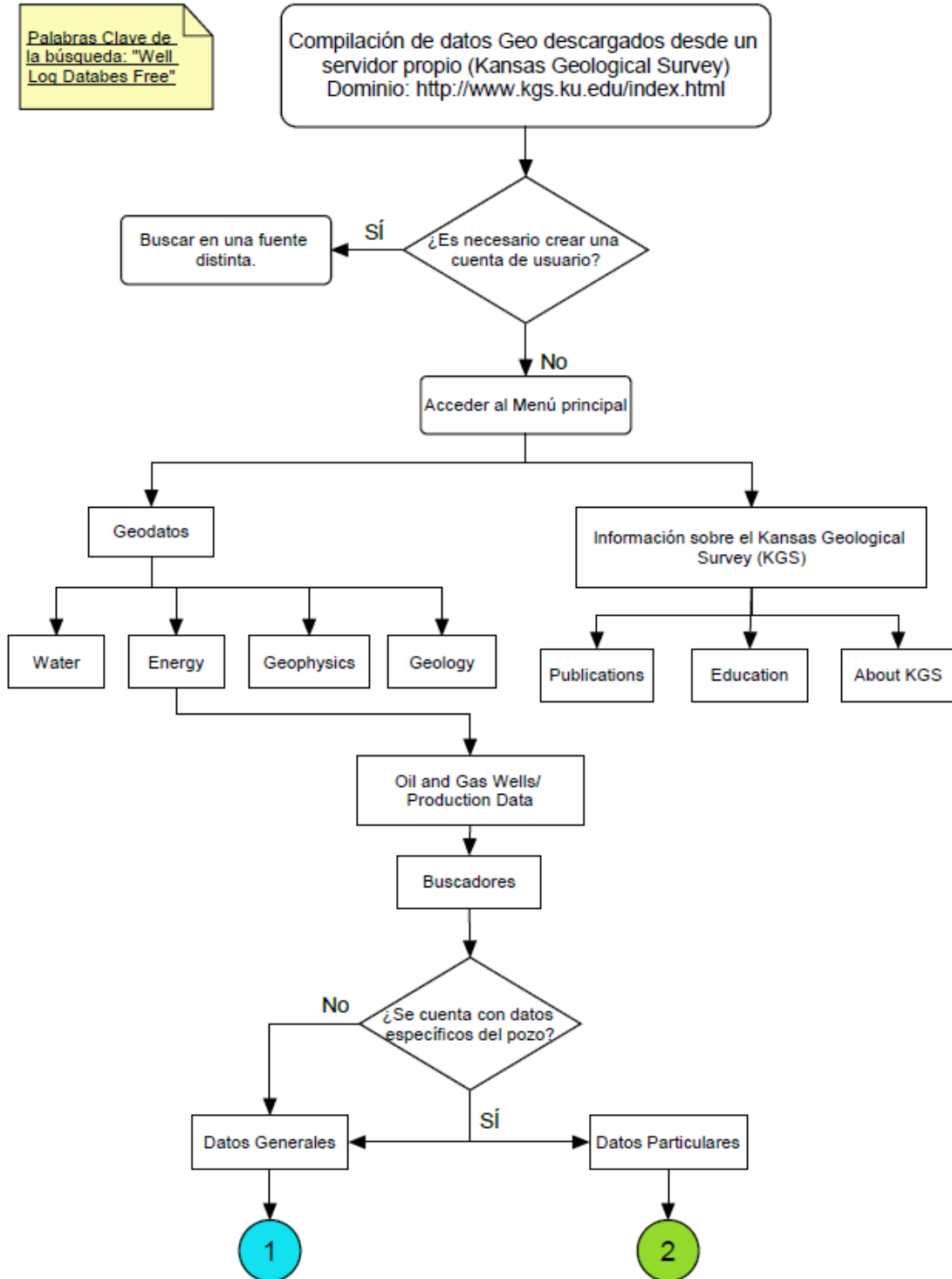
32. Ezekwe, N. (2011). *Petroleum Reservoir Engineering Practice*. Prentice Hall.
33. Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, Methods and Applications*. Berlin: Springer.
34. Fetter, C. (2001). *Applied Hydrogeology* (4a ed.). Prentice-Hall.
35. Fisher, R. (1918). *The Correlation Between Relative on the Supposition of the Mendelian Inheritance*. Raleigh, Carolina del Norte: Department of Genetics North Caroline State Collage.
36. Freeze, A., & Cherry, J. (1979). *Groundwater*. Prentice Hall.
37. García Benitez, S., & et al. (febrero de 2016). Neural Networks for Defining Spatial Variation of Rock Properties in Sparsely Instrumented Media. *Boletín de la Sociedad Geológica Mexicana*, 68(3), 19.
38. Glossa, J. (1982). *Depositional Environments and Diagenetic History of the Nolans Limestone, Rice County, Kansas*. University of Kansas.
39. Goebel, E. (1966). Thermal Recovery Projects are Increasing in Kansas and Missouri. *World Oil*, 162(4), 78-80.
40. Goodman, L., & Kruskal, W. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, 49(268), 732-764.
41. Grus, J. (2015). *Data Science from Scratch*. CA: O'Reilly Media.
42. Güller, B., & Ertekin, T. (1999). An Artificial Neural Network Based Relative Permeability Predictor. *Petroleum Conference of the South Saskatchewan* (págs. 1-25). Regina: Petroleum Society of CIM.
43. Hallsworth, C., & Knox, R. (1999). *Rock Classification, sediments and sedimentary rocks* (Vol. 3er). Nottingham, UK: British Geological Survey.
44. Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. MIT Press.
45. Hernández Ambrosio, I. (2017). *Aplicación de Redes Neuronales en la Ingeniería Petrolera*. Ciudad de México, México: Universidad Autónoma de México.
46. Hilera, J., & Martínez, V. (1995). *Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones* (primera ed.). Madrid: RA-MA Editorial.
47. Honarpour, M., Djabbarah, N., & et al. (2003). Whole Core Analysis - Experience and Challenges. *Whole Core Analysis - Experience and Challenges* (pág. 16). Baharain: SPE.
48. Horner, D. (1951). Pressure Build-Up in Wells. *Third World Petroleum Congress* (pág. 19). Leiden: E. J. Brill.
49. Hosmer, D., & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley & Sons.

50. Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(6), 417-441.
51. Huang, Z., Shumeld, J., & et al. (febrero de 1996). Permeability Prediction with Artificial Neural Network Modeling in the Venture Gas Field, Offshore Eastern Canada. *Gepophysics*, 15.
52. Kaiser. (1974). An Index of Factor Simplicity. *Psychometrika*, 39, 31-36.
53. KGS. (12 de 08 de 2020). *KU: Kansas Geological Survey - The University of Kansas*. Obtenido de <http://www.kgs.ku.edu/PRS/petroIndex.html>
54. Klinkenberg, L. (1941). *The Permeability of Porous Media to Liquids and Gases*. Drill and Production Prac.
55. Kohli, A., & Arora, P. (2014). Application of Artificial Neural Networks for Well Logs. *International Petroleum Technology Conference* (págs. 1-8). Doha: IPTC.
56. Kurzweil, R. (1990). *The Age of Intelligent Machines*. MIT Press.
57. Lee, J. (1982). *Well Testing* (Vol. 1). SPE Textbook.
58. Luger, G. F. (1995). *Computation and Intelligence: Collected Readings*. AAI Press.
59. Maslennikova, Y. (2013). Permeability Prediction Using Hybrid Neural Network Modeling. *SPE Annual Technical Conference* (págs. 1-6). New Orleans: SPE.
60. Maher C. and Collins B. (1949). Hugoton Embayment of Anadarko Basin in Southwestern Kansas, Southeastern Colorado, and Oklahoma Panhandle: Geological Notes. *AAPG Bulletin*, 32(5), 813-816
61. Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Universidad Tecnológica Nacional-Departamento de Ingeniería Química.
62. Matthews, C., & Russel, D. (1967). *Pressure Buildup and Flow Test in Wells* (Vol. 1). Dallas: SPE.
63. McCain, W. (1990). *The Properties of Petroleum Fluids* (Segunda ed.). Tulsa: PennWell Books.
64. Mohaghegh, S. (Septiembre de 1997). Permeability Determination from Well Log Data. *SPE formation evaluation*, 170-177.
65. Monicard, R. (1980). *Properties of Reservoir Rocks: Core Analysis*. Paris: Edition Technip.
66. Muskat, M. (1937). *The Flow of Homogeneous Fluids Through Porous Media*. New York: McGraw-Hill.
67. Naranjo Agudelo, A. (2009). *Evaluación de Yacimientos de Hidrocarburos* (Primera ed.). Medellín: Centro de Publicaciones, Universidad Nacional de Colombia.

68. Nilsson, N. (1998). *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann.
69. North, D. (1968). A Tutorial Introduction to Decision Theory. *Systems Science and Cybernetics*, 3(4).
70. Pape, H., Clauser, C., & Iffland, J. (septiembre de 1999). Permeability prediction based on fractal pore-space geometry. *Geophysics*, 64(5), 1447-1460.
71. Parrado, A. (2016). *Recomendaciones para Manejo y Preservación de Núcleos para Entrega a Litoteca del SGC*. Colombia: Servicio Geológico Colombiano.
72. Pérez Pacheco, G. (2011). *Influencia de Parámetros Petrofísicos en la Determinación Indirecta de la Permeabilidad Absoluta en Rocas de Yacimientos Petroleros*. Ciudad de México: Universidad Nacional Autónoma de México.
73. Piatetski, G., & Frawley, W. (1991). *Knowledge Discovery in Databases*. Cambridge: MIT Press.
74. Pillado Torres, M. (2016). *Tecnologías para el Corte, Manipulación, Preservación y Análisis de Núcleos*. Ciudad de México: Universidad Nacional Autónoma de México.
75. Pinder, G., & Celia, M. (2006). *Subsurface Hydrogeology*. New Jersey: John Wiley & Sons, Inc.
76. Ponce Gallegos, J., Torres Soto, A., Quezada Aguilera, F., & et al. (2014). *Inteligencia Artificial* (1a ed.). Proyecto LATIn.
77. Press, S., & Wilson, S. (1978). Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, 73(364), 699-705.
78. Puckette, J. (1996). The Upper Morrow Reservoirs: Complex Fluvio-deltaic Depositional Systems. *Oklahoma Geological Survey Circular*(98), 47-84.
79. Quinlan, R. (1992). Learning with Continuous Classes. *5th Australian Joint Conference on Artificial Intelligence*, (págs. 343-348). Singapore.
80. Ranjan, S. (2016). *Data Science: Theories, Models, Algorithms and Analytics*. R.D.
81. Rascoe, B., & Adler, F. (1983). Permo-Carboniferous Hydrocarbon Accumulations, Midcontinent. *USA: AAPG Bulletin*, 67(6), 979-1001.
82. Riaño Rupilanchas, D. (2017). On the Origin of Karl Pearson's Term "Histogram". *Revista Estadística Española*, 29-35.
83. Russell, S., & Norvig, P. (2010). *Artificial Intelligence a Modern Approach*. Prentice Hall Series.
84. Sánchez San Roman, J. (2005). Ley de Darcy. Conductividad Hidráulica. *Depto de Geología*, 14.

85. Selley, R. (1998). *Elements of Petroleum Geology* (Segunda ed.). San Diego: Academia Press.
86. Serna Pineda, S. (2009). *Comparación de Árboles de Regresión y Clasificación y Regresión Logística*. Medellín, Colombia: Universidad Nacional de Colombia.
87. Shang, B., Hamman, J., & et al. (5-8 de Octubre de 2003). A Model to Correlate Permeability with Efficient Porosity and Irreducible Water Saturation. *SPE*, 8.
88. Shunway, R., & Stoffer, D. (2010). *Time Series Analysis and Its Applications: with R examples*. Berlin: Springer.
89. Singh, S. (2005). Permeability Prediction Using Artificial Neural Network (ANN): A Case Study of Uinta Basin. *SPE Annual Technical Conference* (pág. 8). Dallas: SPE.
90. Spearman, C. (1904). "General Intelligence": Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201-292.
91. Srayne, K., Ravi, K., & Sergei, V. (?). Cross-Validation and Mean-Square Stability. *Yahoo! Research*, 10.
92. Timofeev, R. (2004). *Classification and Regression Trees (cart). Theory and Applications*. Berlin: Humboldt University.
93. Timur, A. (1968). An Investigation of Permeability, Porosity, and Residual Water Saturation Relationship for Sandstone Reservoirs. *The Log-Analyst*, ?
94. Tixier, M. (1949). *Evaluation of Permeability from Electric-Log Resistivity Gradients*. Oil & Gas J.
95. Torsaeter, O., & Abtahi, M. (2003). *Experimental Reservoir Engineering - Laboratory Workbook*. Noruega: Department of Petroleum Engineering and Applied Geophysics .
96. Tucker, M. (1991). *Sedimentary Petrology*. Backwell Science.
97. Tukey, J. (1977). *Exploratory Data Analysis*. Londres: Pearson.
98. Varhaug, M., & Smithson, T. (mayo de 2015). The Defining Series: Downhole Coring. *Oilfield Review*, 27(1), 4.
99. Vatulkin, I., & Weihs, C. (2017). *Music Data Analysis - Foundations and Applications*. Boca Raton: CRP Press.
100. Weihs, C., & Ickstadt, K. (2018). Data Science: the Impact of Statistics. *International Journal of Data Science and Analytics*, 6.
101. Winston, P. H. (1992). *Artificial Intelligence* (Vol. 3°). Addison-Wesley.

Apéndice A



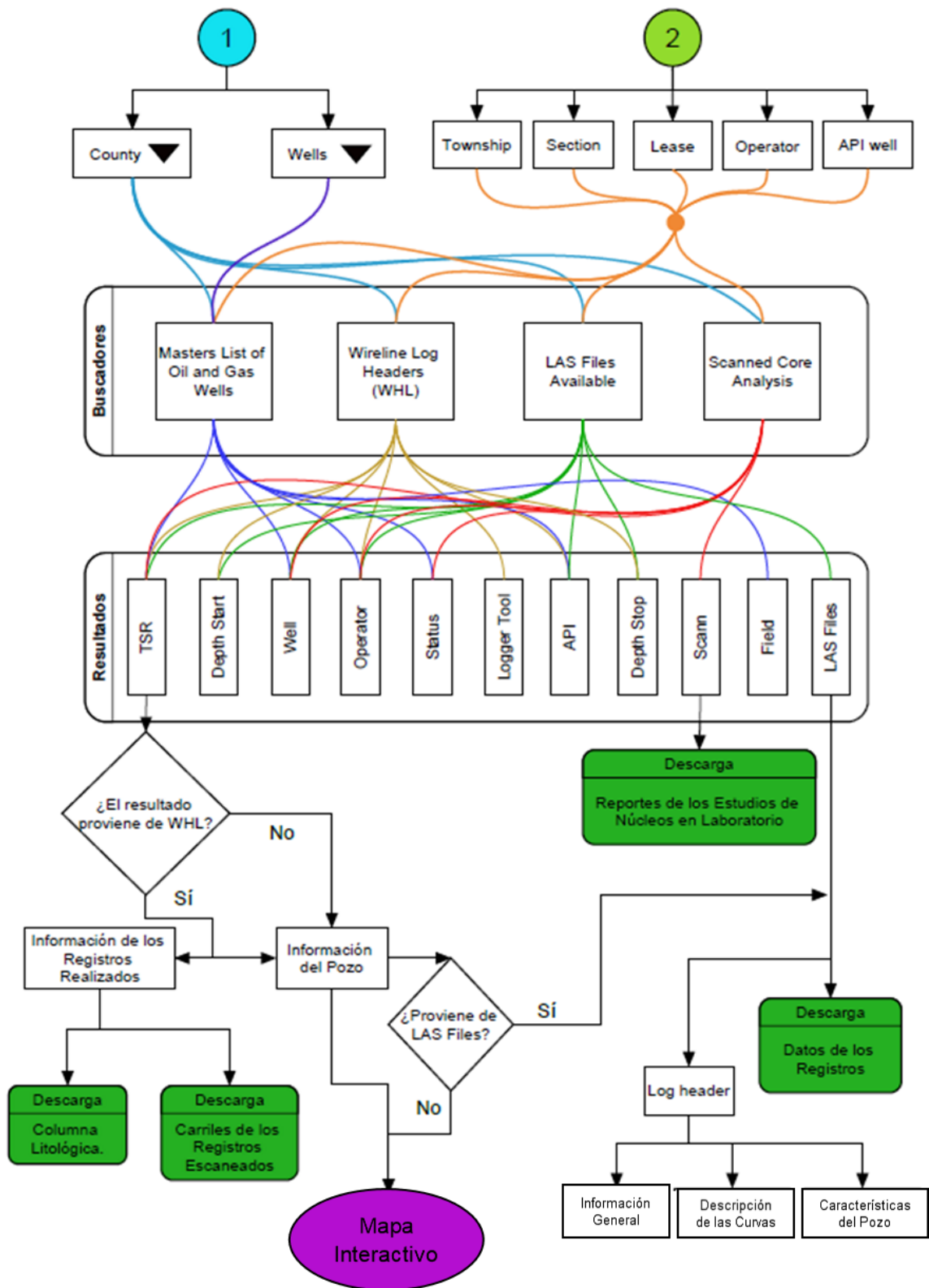


Figura A.1. Diagrama de bloques del repositorio en línea de la KGS.