UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

PROGRAMA DE MAESTRIA Y DOCTORADO EN INGENIERIA

FACULTAD DE INGENIERIA

METODO DE VALIDACION DEL DATO DE PRODUCCION APLICANDO TECNICAS DE MINERIA DE DATOS

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN INGENIERIA

PETROLERA Y DE GAS NATURAL - YACIMIENTOS

P R E S E N T A :

OLIVIA PATRICIA QUIÑONEZ GAMEZ

TUTOR:
DR. RODOLFO G. CAMACHO VELAZQUEZ

2010

JURADO ASIGNADO:

Presidente:	Dr. Fernando Samaniego Verduzco
Secretario:	Dr. Guillermo C. Domínguez Vargas
Vocal:	Dr. Rodolfo G. Camacho Velázquez
1er. Suplente:	Dr. José Antonio González Guevara
2do. Suplente:	M. en I. Mario A. Vásquez Cruz

Lugar o lugares donde se realizó la tesis:

México, Distrito Federal.

TUTOR DE TESIS:

Dr. Rodolfo G. Camacho Velázquez

FIRMA

DEDICATORIA

A mis padres, Juan y Monchi, porque siempre han sido una fuente de inspiración para alcanzar mis metas.

A mis hijos Andrea Carolina, Juan Luis y Manuel Alejandro, por ser el motor de mi vida, por el tiempo que les robé por alcanzar esta meta y porque no olvido sus palabras: “Mamá, si tu eres feliz nosotros vamos a ser felices”.

A mi amado esposo Iván, por su apoyo, paciencia, comprensión y sobre todo, por su gran amor.

A mis hermanos, Claudia, Rocío, Sergio y Juan, quienes siempre me animaron a continuar.

A Rosita, por ser parte de nuestra vida.

AGRADECIMIENTOS

A Dios, por darme tanto.

A mis sinodales: Dr. Fernando Samaniego Verduzco (Presidente), Dr. Guillermo Domínguez Vargas (Secretario), Dr. Rodolfo Camacho Velázquez (Vocal), Dr. José Antonio González Guevara (Primer Suplente) y M.I. Mario Vásquez Cruz (Segundo Suplente), por sus comentarios y la revisión final de este trabajo.

De manera muy especial a mi tutor, el Dr. Rodolfo G. Camacho Velázquez, por su apoyo, su valioso aporte y por el tiempo dedicado a la asesoría de este trabajo.

A mis maestros de la Facultad de Ingeniería, especialmente al Dr. Fernando Samaniego y al Dr. Guillermo Domínguez, y a todas las personas que hicieron posible la culminación de este trabajo.

A mis profesoras Teo y Nachita, quienes sembraron en mi la semilla del conocimiento, y a quienes recuerdo con mucho cariño.

A todos mis compañer@s y amig@s quienes compartieron conmigo sus conocimientos y experiencias en las aulas. Un reconocimiento a Elsa por su amistad y gran apoyo. A mis amig@s, l@s que no estuvieron en las aulas, pero siempre creyeron que lo lograría. A Pilita, porque estuvo cuando más la necesité.

A mi Alma Máter, la Universidad Nacional Autónoma de México, por brindarme nuevamente la oportunidad de crecer.

A PEMEX Exploración y Producción, porque permitió que esto fuera posible.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

La calidad de los datos de producción es un tema de interés general en la industria petrolera. En muchas ocasiones es la única información disponible en cantidad suficiente en los campos maduros. Gracias al desarrollo de la tecnología de información los datos digitalizados son altamente disponibles; sin embargo, no todos son confiables, lo que puede causar fallas operacionales y conducir a la toma inadecuada de decisiones.

Con el fin de determinar la calidad de los datos de producción de un campo petrolero, se desarrolló una metodología, la cual se apoyó en técnicas de Minería de Datos. Esta metodología combinó un algoritmo de clasificación difusa, modelado de redes neuronales y un proceso iterativo. Se aplicó a un caso real: una base de datos de un campo petrolero marino. El resultado fue la clasificación de los datos en: buenos, ligeramente contaminados o malos. Posteriormente, se evaluó el comportamiento de la declinación de un pozo. El resultado obtenido con los datos buenos y ligeramente contaminados fue el esperado para un pozo.

Se pudo concluir que esta metodología de clasificación generó una solución simple para el problema de calidad de los datos de producción. La metodología desarrollada puede aplicarse a cualquier conjunto de datos. Los resultados infieren que existe un grado de subjetividad en la metodología, ya que al cambiar los criterios de restricción de clasificación, la calidad del dato resultó diferente. Adicionalmente, durante la aplicación de la metodología, pudimos comprobar la eficacia de las herramientas de Minería de Datos en la estimación de datos faltantes.

CAPÍTULO 1. INTRODUCCIÓN

El éxito en la industria petrolera descansa en la combinación de información multidimensional confiable y técnicas interdisciplinarias. La búsqueda de información confiable requiere de más atención que la que actualmente se le dedica, ya que es frecuente encontrar que no hay certeza acerca de la calidad de los datos obtenidos de los diferentes procesos petroleros. Como se sabe, la existencia de datos contaminados puede causar fallas operacionales y además, conducir a la toma de decisiones inadecuada a nivel estratégico.

En otras áreas diferentes a la industria petrolera, se han presentado problemas similares que se han analizado y resuelto con técnicas poco convencionales, pero con muy buenos resultados. Estas técnicas están basadas en los procesos elementales de la evolución. Recordemos que la evolución es un principio del pensamiento moderno en la biología. La teoría de la evolución clásica de Darwin, combinada con la de selección de Weissman, así como la genética de Mendel, son consideradas como el paradigma neo-Darwiniano. El neo-Darwinismo sugiere que la vida puede explicarse mediante unos pocos procesos estadísticos actuando sobre poblaciones. Estos procesos son la reproducción, la mutación, la competencia y la selección. La evolución es el resultado de la interacción entre estos procesos. El pensamiento evolutivo se extiende más allá del estudio de la vida, ya que al considerar a la evolución como un proceso de optimización, éste se puede aplicar a problemas de ingeniería, que pueden simularse usando una computadora. El interés en tales simulaciones se ha incrementado con el desarrollo de aplicaciones para reemplazar tecnologías convencionales en sistemas de control automático, de reconocimiento de patrones, sistemas financieros, etc.

Desde hace tiempo, los modelos matemáticos analíticos han sido el centro de todas las ciencias y también han modelado la mayoría de los principios de ingeniería. Sin embargo, los modelos analíticos están limitados a las condiciones en las que se encontró la solución matemática. Sobre todo las soluciones matemáticas que describen procesos como el de flujo de fluidos, que deben resolverse simultáneamente con las

soluciones que describen procesos como la transferencia de calor o la resistencia de materiales. Es decir, la solución de estos problemas no puede obtenerse en forma independiente porque frecuentemente se ven afectados unos a otros. Por ejemplo, la presión del flujo de fluidos puede afectar la tensión y en consecuencia la resistencia del material. Algunas de estas dificultades han disminuido con los grandes avances en matemáticas aplicadas y métodos numéricos. La simulación numérica, que ha sido beneficiada en forma importante al mejorar el desempeño de las computadoras, es una de las que más ha contribuido. Aún así, el poder de los nuevos métodos computacionales no puede sustituir la lógica en la definición del problema, ni la definición de todos los fenómenos que lo influencian. Si sobreestimamos o ignoramos la participación de algunos fenómenos se pueden generar fallas en el diseño y obtener un pronóstico incorrecto del comportamiento futuro.

La habilidad para predecir el comportamiento futuro es el centro de todas las ingenierías. Gracias a los avances importantes en medición e instrumentación, se han generado grandes volúmenes de datos, mucho más rápido que el tiempo requerido para interpretarlos e incluso almacenarlos. Además, muchos procesos pueden resultar en datos que son difíciles de interpretar, ya sea porque no se ajustan al comportamiento esperado o porque se ignora alguna componente importante. Posiblemente un gran número de fenómenos pueden estar involucrados y, consecuentemente, existir alguna componente cuya influencia puede estar oculta o inconsistente.

El proceso de descubrir nuevas correlaciones significativas, patrones y tendencias, filtrando grandes cantidades de datos mediante el uso de tecnologías de reconocimiento de patrones, así como de técnicas matemáticas y estadísticas, ha sido denominado *Minería de Datos.*

Una de las metas de esta tesis es proporcionar un primer acercamiento a este proceso y aplicarlo a la industria petrolera en México. Es importante señalar que la minería de datos no ignora la física de los fenómenos y procesos, sino que reconoce la diversidad y complejidad de sus influencias, y permite entender esta complejidad.

La minería de datos es la herramienta principal para la obtención de conocimiento a partir de los datos.

La industria petrolera se ha familiarizado con este conjunto de herramientas desde principios de los 90's, las cuales se han usado en numerosas aplicaciones a lo largo de la cadena productiva, incluyendo la interpretación de registros de pozos hasta la optimización de fracturamientos hidráulicos.

La vida diaria de los profesionales del petróleo está llena de problemas dinámicos altamente complejos, situación que actualmente se está viviendo en la industria petrolera, por lo que se pueden obtener beneficio de lo que la minería de datos puede ofrecer.

En las últimas dos décadas las compañías petroleras han gastado millones de dólares para almacenar datos digitales o para convertir los datos existentes a formato digital. Esto se debe a que se han dado cuenta del valor de los datos y el poder que estos tienen en la mejora de sus operaciones. Actualmente, la mayor parte de las compañías petroleras cuentan con grandes bases de datos de sus campos de gas y aceite, que contienen información relacionada con la producción de hidrocarburos. Sin embargo, no todos los registros de datos son completamente confiables o simplemente, no reflejan la realidad. Los errores en los datos almacenados pueden ser subjetivos u objetivos, y ser el resultado de una captura de datos inapropiada o incompleta, calibración inadecuada o fallas en los equipos de medición, una mala interpretación o algún otro motivo. Estos errores pueden llevarnos a una deficiente, errónea o hasta una interpretación imposible de los datos, lo cual conduce a cuestionarnos acerca de cuántos datos son realmente confiables y cómo podemos identificar los datos contaminados.

Esta tesis presenta una metodología nueva para identificar datos contaminados en las bases de datos de producción. Para lograr su objetivo, esta metodología se apoyará en las técnicas de minería de datos, de las cuales utilizaremos herramientas basadas en algoritmos de clasificación difusa y de redes neuronales artificiales. El resultado de utilizar esta metodología podrá considerarse como una clasificación de los datos en: malos, ligeramente contaminados o buenos.

Para el desarrollo de este trabajo, se inicia con una revisión de los casos documentados en la literatura relacionados con la aplicación de las técnicas de minería de datos en la industria petrolera, de los cuales rescataremos los aspectos que sean de interés y

ayuden a lograr los objetivos de esta tesis; enseguida se presentarán los fundamentos del análisis de la declinación de la producción. Posteriormente se documentará la teoría relacionada con las técnicas de minería de datos que sustentan este trabajo, que como se mencionó previamente son los algoritmos de clasificación difusa y de redes neuronales artificiales. Después, tomando como base tanto la teoría del análisis de declinación de la producción como las técnicas de minería de datos, se desarrollará la metodología que permitirá validar los datos de producción. Para comprobar la metodología propuesta, se presentará un caso real de aplicación. Finalmente, se presentará el análisis de los resultados con sus conclusiones.

CAPÍTULO 2. REVISIÓN DE LITERATURA

Este capítulo presenta un panorama general del estado en que se encuentra la aplicación de las técnicas de minería de datos en la industria petrolera.

Durante la búsqueda de literatura relacionada con el tema encontramos que, aunque existen diversas aplicaciones de la minería de datos en la industria petrolera, no se tienen suficientes casos, o al menos documentados, en los que se haya aplicado a problemas de validación de datos. Sin embargo, fue posible rescatar algunos datos interesantes, así como identificar la metodología que se ha utilizado en la solución de problemas de ingeniería petrolera que han requerido una validación de datos previa. Es importante señalar que en todos los casos documentados, siempre se menciona esta validación de datos como parte de una metodología general; asimismo, se ha observado que una herramienta integrada de minería de datos debe contemplar como componente imprescindible un módulo de importación, limpieza y agrupamiento de datos. Por otro lado, el pre-procesamiento o preparación de datos, es una de los componentes más importantes del proceso de minería de datosi.

Mohagheghii, quien ha sido uno de los promotores principales de la aplicación de los sistemas inteligentes a la solución de problemas en la industria petrolera, comenta que una herramienta de software inteligente debe tener una serie de atributos importantes, tales como la habilidad para integrar la parte "dura" (estadística) con la parte "suave" (inteligencia) de la computación, y conjuntar diferentes técnicas de inteligencia artificial (lógica difusa, redes neuronales, optimización genética y motores de inferencia difusa). Cualquier herramienta con las características descritas, deberá dirigirse a profesionales de la industria petrolera con una transparencia que permita eliminar su imagen de "caja negra".

El mismo investigador[ii] indica que los sistemas inteligentes pueden usarse para tratar muchos problemas encontrados en nuestra industria, los cuales pueden dividirse en las categorías siguientes:

1. Activado o conducido totalmente por datos. Este tipo de problemas usan una gran cantidad de datos con el fin de modelar la dinámica de un sistema. Esto es

principalmente un enfoque empírico, ya que no está basado en interpretar las dependencias o leyes de las ciencias que describen el fenómeno físico. Problemas tales como el desarrollo de registros sintéticos, caracterización de yacimientos por correlación de registros a sísmica y datos de núcleo, así como la predicción de la producción de gas natural, son algunos ejemplos.

2. Basado totalmente en reglas. Estos son problemas de toma de decisiones, en los cuales debe usarse el conocimiento de un experto. La Interpretación de registros de pozo y la identificación de métodos de recuperación mejorada son buenos ejemplos.

3. Optimización. Estos problemas están relacionados con encontrar el mejor conjunto de condiciones/operaciones para lograr un resultado específico en problemas dinámicos, no lineales y altamente complejos. Ejemplos de estos son la optimización de instalaciones superficiales para incrementar la producción de aceite.

4. **Fusión de conocimiento y datos.** Estos son problemas que integran datos con conocimiento experto para abordar temas complejos, tales como la selección de pozos candidatos para estimulación y la identificación de mejores prácticas.

Con base en la clasificación propuesta por Mohaghegh[ii] el problema de validación del dato de producción entra en la categoría de problemas activados o conducidos totalmente por datos.

En el mismo artículo también se menciona que la imaginación del ingeniero es lo único que limita la aplicación de este tipo de herramientas en la industria petrolera.

La tarea principal del profesionista petrolero es identificar qué tipo de problemas se benefician mayormente por los sistemas inteligentes. Un sistema inteligente integrado, tal como cualquier otra tecnología, no va a ser la panacea de nuestra industria, pero puede desempeñar un papel muy importante para llevarlo hacia las fronteras de la tecnología de información.

Otro comentario de interés basado en la experiencia del autor, indica que el desarrollo exitoso de un modelo de red neuronal requiere la integración de lógica difusa y optimización genética. Estos conceptos se revisarán con detalle en el capítulo cuatro.

Al igual que cualquier sistema analítico, los sistemas inteligentes tienen limitaciones. Es importante entender las limitaciones de estas técnicas para incrementar la probabilidad de éxito y eficiencia. En el caso de las redes neuronales la limitación es la insuficiencia de datos. Por lo tanto es muy importante considerarla en su aplicación.

Popa y colaboradoresiii, presentan una metodología para identificar los datos contaminados en bases de datos de fracturamiento hidráulico. La metodología combina una serie de herramientas de inteligencia artificial, las cuales integran técnicas de clasificación difusa, modelado de redes neuronales y un proceso iterativo para lograr su meta (identificar y eliminar datos contaminados en la base de datos para continuar posteriormente con la identificación de mejores prácticas).

La metodología propuesta por Popa y colaboradores[iii] contempla los puntos discutidos a continuación:

Datos de entrada

Consiste en una serie de etapas: en la primera se seleccionan los parámetros de mayor influencia en el problema a resolver, que para su ejemplo fueron los parámetros relacionados con las estimulaciones y el fracturamiento. Después, se seleccionan los parámetros de salida del sistema, que para este caso fue el valor real de la producción máxima obtenida después del fracturamiento.

Control de calidad de los datos

Para el ejemplo en cuestión el primer paso consistió en identificar los datos fuera de la tendencia; es decir aquellos cuyo comportamiento fuese muy diferente al resto. Estos datos causaban que la salida fuera cinco veces superior a la producción pico. Por tanto, después de investigar el origen de estos datos y entender las causas de su comportamiento anómalo, fueron removidos y excluidos del análisis.

Dentro del control de calidad también se generaron gráficas de regresión (un parámetro contra otro para detectar cualquier tendencia visual entre ellos) y gráficas de distribución de frecuencias. En este caso no se obtuvieron correlaciones visibles.

Aún con la limpieza de la base de datos, los resultados de la red neuronal no fueron exitosos, por lo que se concluyó que los datos estaban contaminados. Ahora el

problema era identificarlos y eliminarlos. Esto se resolvió a través del sistema de clasificación de datos descrito a continuación.

Sistema de clasificación de datos Neuro-Cluster

La metodología usa un conjunto de herramientas de inteligencia artificial, las cuales incluyen técnicas de clasificación basada en la teoría de lógica difusa, modelado de redes neuronales artificiales y un proceso iterativo, para alcanzar una meta convergente. El resultado es la clasificación de los datos.

En este sistema se usan los parámetros de mayor influencia. Primero, se clasifican los datos. La salida del sistema se incluye dentro del conjunto de datos antes de ser clasificado. Los datos se clasifican en grupos difusos. La información de clasificación junto con el grado de caos o entropía de cada elemento, se agrega al conjunto de datos y una red neuronal artificial es entrenada con un desempeño relativamente bueno**.** Con el fin de concluir cuando un dato es bueno, ligeramente contaminado o malo, se establecen criterios con base en tres curvas, las cuales identifican la información de clase, la entropía, y la diferencia entre el valor actual y el valor estimado por la red neuronal. Cada punto en la curva es un valor discreto de la salida. La clave en este proceso es la presencia de datos de salida en el proceso de clasificación.

En el artículo en cuestión se describe de manera general el algoritmo utilizado para la clasificación, el cual fue muy importante para el desarrollo de esta tesis.

La metodología aplicada a datos de fracturamiento hidráulico mostró que si se aplica cuidadosamente a una base de datos que contenga registros corruptos, estos pueden identificarse exitosamente. De esta forma, la aplicación de la metodología puede extenderse a cualquier tipo de bases de datos para la identificación de información contaminada.

La combinación de estas dos herramientas inteligentes, redes neuronales artificiales y algoritmos de clasificación difusa, es innovadora y proporciona una solución simple a problemas de validación de datos.

Otro punto de interés para el trabajo presente se encontró en una publicación de Hernándeziv, donde nos presenta su punto de vista acerca de las fases de un proyecto

de minería de datos. De acuerdo a su experiencia, un proyecto de minería de datos debe incluir las fases siguientes.

Filtrado de Datos

Se determinan las fuentes de información que pueden ser útiles, así como el formato a utilizarse. El formato de los datos contenido en la fuente de información casi nunca es el idóneo, y la mayoría de las veces no es posible utilizar ningún algoritmo de minería de datos. Mediante el pre-procesado se filtran los datos (se eliminan valores incorrectos, no válidos y desconocidos), se obtienen muestras de los mismos, o se reducen el número de valores posibles (mediante redondeo).

Selección de variables

Aún cuando se hayan filtrado los datos, su cantidad sigue siendo enorme, por lo que se procede a la selección de solo algunas características para reducir su número. Para lograr lo anterior se eligen las variables que más influyen en el problema. Los métodos para la selección de características son dos: el primero se basa en la selección de los atributos que mejor describen el problema y el otro basado en variables independientes mediante pruebas de sensibilidad, algoritmos de distancia o heurísticos.

Extracción de conocimiento

Mediante alguna técnica, redes neuronales, algoritmos genéticos y/o algoritmos de lógica difusa, se obtiene un modelo de conocimiento que representa los patrones de comportamiento observados en los valores de las variables del problema, o relaciones de asociación entre dichas variables, aunque también pueden usarse varias técnicas a la vez para generar varios modelos.

Interpretación y evaluación

La última fase consiste en la validación del modelo, para la cual se debe comprobar que las conclusiones sean válidas y satisfactorias. En el caso de haberse obtenido varios modelos mediante el uso de distintas técnicas, deben compararse para elegir el que mejor solucione el problema. Si ninguno de los modelos alcanza los resultados esperados, habrá que modificar algunas de las fases anteriores en busca de nuevos modelos.

Zanglv, presenta ejemplos de cómo los modelos predictivos de minería de datos pueden usarse para mejorar y acelerar la producción de yacimientos de hidrocarburos. Estos modelos actúan como enlace, atendiendo tanto la calidad de los datos, como los cambios de comportamiento de pozos o de sistemas de producción completos.

Además, Zangl[v] menciona que debido al incremento exponencial de la cantidad de datos, actualmente, la industria petrolera está teniendo problemas para analizar y optimizar sus procesos. Paradójicamente, en tiempos de alta tecnología, la calidad de los datos y de la información derivada de éstos, va en descenso.

Por otro lado, Zanglvi documenta el uso de las herramientas de minería de datos en la industria petrolera. Menciona que muchos de los nuevos métodos computacionales que no se consideran en la práctica cotidiana, muestran resultados sorprendentes. Por ejemplo, las redes neuronales artificiales tienen muchas aplicaciones en el modelado y predicción del comportamiento de un yacimiento, en la estimación de gastos, presiones, cortes de agua, predicción de permeabilidades a partir de registros de pozos, en procesos de optimización y en métodos de selección de pozos candidatos para procesos de estimulación.

Algo que impide la aceptación de estos métodos en la industria, es que la mayoría de los ingenieros continúan sintiéndose más cómodos usando métodos tradicionales.

La diferencia más importante entre los métodos tradicionales y los nuevos métodos es que los primeros se orientan hacia el conocimiento; los nuevos métodos son principalmente activados o conducidos por los datos.

El análisis activado o conducido por conocimiento entrega respuestas a problemas de ingeniería a través de la aplicación de ecuaciones matemáticas basadas en leyes científicas. El programador y el usuario de un programa de cómputo de este tipo, tiene que estar familiarizado con las limitaciones al aplicar estas leyes. Así, en aplicaciones más complejas, el programador debe definir las suposiciones y simplificaciones que se emplearon para desarrollar el método de solución. Una capacitación especializada del ingeniero es necesaria para obtener resultados confiables.

Un buen ejemplo es la simulación numérica de yacimientos. Esta área ha logrado una posición estratégica dentro de la industria petrolera. Todos los simuladores comerciales

disponibles están basados en conceptos matemáticos similares y por lo tanto usan suposiciones y limitaciones similares. No obstante, cada uno tiene un diseño ligeramente diferente y ofrece funcionalidad diferente. Como resultado, ninguno produce exactamente la misma salida aunque usen la misma entrada. Si el usuario ignora o no considera estas diferencias, será incapaz de entender y de interpretar correctamente los resultados del programa.

El análisis activado o conducido por datos supone que los algoritmos matemáticos encuentran relaciones, tales como patrones recurrentes o clases en los datos. El análisis y el sistema que se emplea, no están basados en las interpretaciones de dependencias o leyes que han sido descritas por las ciencias naturales. Operar este tipo de programas de cómputo requiere conocimientos básicos de computación para manejar el flujo de datos y desplegar los resultados. Sin embargo, los datos deben prepararse con cuidado. La teoría subyacente en los programas de minería de datos está basada en ecuaciones algebraicas. La combinación de estos métodos y de computadoras con alto desempeño, permite el análisis de grandes cantidades de datos en miles de ciclos iterativos.

Zangl[vi] clasifica las tareas de minería de datos en las categorías siguientes:

Clasificación

Esta tarea consiste en etiquetar los registros de datos. Un ejemplo típico es la delineación de litofacies de datos de registros de pozos, la cual puede ser una tarea demandante y que tiene que ser repetida para cada uno de los pozos. La misma tarea puede resolverse usando mapas auto-organizados (SOM - Self Organized Maps), una clase de red neuronal.

Estimación

Es la tarea que consiste en llenar valores faltantes en un campo especial de un registro del conjunto de datos como una función de los otros registros del conjunto de datos. Las técnicas de regresión usuales se emplean frecuentemente para la estimación, que es también una aplicación común de una red neuronal artificial. Por ejemplo, llenar los espacios vacíos en la historia de producción y presión es una tarea de estimación. Una red neuronal puede usarse para estimar volúmenes de producción diaria de un pozo,

basándose en las dependencias aprendidas en pruebas de producción. Los parámetros de una prueba del pozo tales como presión de cabeza, volumen de gas, etc., se usan como parámetros de entrada mientras la producción de aceite medida es la salida. La caja negra, que es la red neuronal, encuentra las relaciones entre entradas y salidas usando un conjunto de datos "entrenado". Una vez que el entrenamiento ha convergido, los parámetros de la prueba del pozo se sustituyen por valores medidos diariamente. Alimentando la red neuronal con datos diarios y no esporádicos, permite obtener valores diarios de salida.

Segmentación

La segmentación consiste en dividir la población total de datos en pequeñas sub-poblaciones que tienen comportamientos similares. Dentro de estas sub-poblaciones se pueden realizar toda clase de predicciones. Por ejemplo, el contenido de agua de varias rocas del yacimiento tiene un papel crucial en la declinación de un yacimiento de gas y aceite. Así, es de interés especial dividir las rocas del yacimiento de acuerdo a su contenido de agua. Las rocas con alto contenido de agua tendrán propiedades totalmente diferentes que las impregnadas de aceite. Los registros de resistividad varían para diferentes niveles de saturación. Entonces, es crucial subdividir rocas de acuerdo a su contenido de agua.

Descripción

Esta tarea se usa para ejecutar tareas de clasificación cuando las clases por sí mismas no están bien definidas. Un ejemplo de aplicación, donde el reconocimiento de relaciones es algo complicado, es la evaluación de tratamientos de estimulación de pozos. Muchos parámetros influyen en el éxito de los tratamientos y solo una combinación de parámetros clave puede conducir al incremento de producción deseada. Un modelo SOM (Self Organized Maps) puede agrupar los datos en segundos, de tal forma que el ingeniero puede inmediatamente observar los principales parámetros indicadores y los factores de influencia.

Zangl[vi], además de la clasificación anterior, distingue dos grandes categorías de la minería de datos:

Minería de datos descriptiva. Su objetivo es encontrar patrones humanamente interpretables, asociaciones o correlaciones, las cuales describen el comportamiento de los datos.

Minería de datos predictiva, donde los modelos se construyen ejecutando inferencias sobre conjuntos de datos disponibles, e intentando predecir el comportamiento de nuevos conjuntos de datos.

Zangl[vi] usa en su caso de estudio una manera de lograr un nivel más alto de automatización y optimización a través de herramientas de minería de datos predictivas. Muestra el poder de relacionar los modelos de la minería de datos predictiva con un sistema experto, en un ambiente de procesos automatizados.

Además, Zangl[vi] mostró que las herramientas de minería de datos predictiva son una opción poderosa para el control de calidad automatizado, removiendo datos fuera de rango y reemplazando datos faltantes, o valores erróneos, con otros calculados.

Finalmente, otro caso de estudio que no puede ignorarse, es un modelo de red neuronal artificial para la predicción de la presión de fondo fluyendo, $p_{wf,}$ y en consecuencia la caída de presión, presentado por Osmanvii, quien demuestra el poder que tienen estas herramientas para la ingeniería petrolera. Su modelo fue probado con bases de datos de campos del Medio Oriente. Primero se demostró que el modelo es estable y simula el proceso físico. Después compara las predicciones del modelo con las correlaciones existentes. Esto lo hace por medio de gráficas y usan el porcentaje de error promedio como indicador. El nuevo modelo obtuvo predicciones más acertadas que los modelos empíricos y mecanísticos.

CAPÍTULO 3. Fundamentos del análisis de la declinación de datos de producción

La importancia de un análisis e interpretación precisos del comportamiento de un yacimiento empleando solamente la historia de producción y presión en función del tiempo no debe descartarse. En la mayoría de los casos es la única información disponible en cantidad y calidad suficiente en los pozos en etapa avanzada de explotación.

Aun cuando el análisis de la producción para la caracterización del yacimiento está alcanzando la popularidad del análisis del período transitorio de presión, en la práctica hay pocos métodos de diagnóstico consistentes para el análisis de los datos de producciónі. Muchos de los métodos de diagnóstico para el análisis de los datos de producción son prácticamente enfoques basados en observaciones, y son esencialmente “reglas de dedo”.

Aunque los análisis de la presión transitoria del pozo y de datos de producción se basan en las mismas teorías y soluciones, debemos reconocer que los datos transitorios de presión se adquieren como parte de un experimento controlado, desarrollado como un estudio específico (pruebas de incremento de presión, por ejemplo). En contraste, los datos de producción se consideran generalmente como datos de monitoreo, con poco control y variaciones considerables durante la adquisición de estos datos.

Una evaluación elemental podría concluir que los datos registrados del transitorio de la presión son:

- datos de alta resolución
- datos de alta frecuencia,

y por otro lado, que los datos de producción son:

- datos de baja resolución
- datos de baja frecuencia

Pudiera ser un punto de vista muy simplista; sin embargo, es un hecho que los datos de producción rara vez alcanzan la calidad, cantidad o precisión de los datos adquiridos durante un prueba de presión.

Es frecuente que la historia de los datos de producción se obtenga cuidadosamente de los registros diarios (o reportes de ventas), mientras que la historia de presión sea infrecuente, imprecisa, o inexistente.

Independientemente de su calidad, los datos disponibles más comunes, especialmente en el caso de campos maduros, son los datos de producción. Los métodos prácticos para el análisis de datos de producción han recorrido un largo camino desde su introducción en la industria hace varias décadas, y todos ellos caen en dos categorías principales:

- Análisis de Curva de Declinación
- Ajuste de Curva Tipo

Mientras que el Análisis de Curva de Declinación es independiente de algunas características del yacimiento, el Ajuste de Curva Tipo es un procedimiento muy subjetivo.ii

Con el fin de entender los parámetros que afectan el comportamiento de la producción, en este capítulo se mencionarán algunos de los métodos de análisis de datos de producción existentes. Los métodos actuales de análisis de datos de producción pueden proporcionar las características del yacimiento, pero tienen dos grandes inconvenientes.

- Para caracterización del yacimiento, el proceso requiere además del gasto, la presión de fondo o la presión en la cabeza, y la presión promedio del yacimiento. Estas presiones no están disponibles frecuentemente en campos maduros.
- Debido a que estas técnicas están dirigidas a pozos individuales, no existe actualmente una herramienta que permita la integración de resultados de cientos de pozos individuales, para ser estudiados como un campo o yacimiento completo.

Las técnicas de análisis de datos de producción han mejorado significativamente en los últimos años. Estas técnicas se usan para proporcionar información de permeabilidad del yacimiento, longitud y conductividad de fracturas hidráulicas, área de drene del pozo, volumen original de gas, recuperación final esperada y factor de daño. Aunque existen muchos métodos disponibles, no existe uno que siempre proporcione una respuesta confiableiii.

Los métodos de análisis de datos de producción integran gastos y presiones para determinar la recuperación final esperada, volumen original de hidrocarburos, permeabilidad y daño, sin tener que cerrar el pozo.

El método de las curvas de declinación es de tipo estadístico, basado en la recopilación de los datos de producción de un yacimiento petrolero durante su vida productiva fluyente, con la finalidad de evaluar las producciones esperadas.

El Análisis de Curvas de Declinación es un método que permite representar por medio de una función matemática los gastos de producción observados de pozos individuales, grupos de pozos o yacimientos, con el fin de predecir el comportamiento de la producción en el futuro, extrapolando la función de declinación representativa.

3.1. Estado del Arte en el Análisis de Datos de Producción

La literatura existente para el análisis de la producción puede dividirse en las categorías y referencias elementales siguientes [i]:

- **Análisis Básico de Datos de Producción**: El trabajo de Arpsiv fue el primer intento sistemático para correlacionar los datos de producción, por lo que es considerado como un punto esencial del análisis. Mattar y McNeilv desarrollaron un par de teorías de balance de materia y flujo pseudoestacionario las cuales proporcionan una metodología de análisis/interpretación para datos de producción sobre las bases de un solo pozo. Li y Hornevi muestran un intento reciente de formalizar el análisis de producción, proporcionando una base teórica (hasta donde sea posible) para varias de las relaciones más comunes del análisis de la producción.

Blasingame y Rushingvii proporcionaron una sinopsis de los métodos históricamente usados para el análisis simplificado de la producción y nos dan algo de soporte teórico para aplicaciones comunes (por ejemplo, las relaciones de declinación exponencial e hiperbólica, así como una solución semi-analítica para flujo de gas). Camacho y Raghavanviii, proporcionan las bases teóricas para yacimientos con empuje de gas en solución durante el período dominado por la frontera externa, por lo que se consideran como una referencia esencial en el análisis de la producción.

- **Análisis de Declinación con Curvas Tipo**: La referencia obligatoria para el análisis de declinación de la producción usando curvas tipo, es el trabajo original sobre el tema de Fetkovich[xix]. Las bases analíticas y las funciones de graficación integral para datos de producción (gasto/Δp), fueron desarrolladas por Palacio y Blasingameix (para pozos de gas), y por Doublet y Blasingamex para pozos de aceite.

  Estas metodologías de normalización del gasto se desarrollaron para pozos fracturados por Agarwal y colaboradoresxi y Araya y Ozkanxii, quienes señalaron perspectivas importantes sobre el uso del análisis de declinación con curva tipo para pozos verticales, fracturados y horizontales. En el año 2004, Fuentes-C y colaboradoresxiii, propusieron extensiones a los enfoques del análisis de declinación por curva tipo para una variedad de casos en yacimientos naturalmente fracturados, con presencia de vúgulos.

- **Métodos de Diagnóstico para Análisis de Datos de Producción**: Existe poca literatura sobre este tema específico de diagnóstico con respecto al análisis de la producción. En contraste, la literatura cuenta con gran cantidad de estudios sobre el análisis del diagnóstico de datos de pruebas del transitorio de presión. Esta situación existe debido, como se mencionó antes, a que los datos de las pruebas de presión se consideran como datos de alta frecuencia y alta resolución, que pueden indicar un carácter único para una condición particular pozo/yacimiento. Los datos de producción se visualizan como datos de baja frecuencia y baja resolución, por lo que el análisis de datos del gasto está etiquetado por algunos opositores como “análisis nebuloso”. Aunque tal punto de

vista es algo cínico, hay verdad en la percepción de que el diagnóstico de los datos de producción es más arte que ciencia.

Mattar y Andersonxiv,xv presentan los lineamientos y ejemplos para el diagnóstico de los datos de producción respecto al análisis basado en el modelo con curvas tipo. Kabir e Izgecxvi proveen una guía sobre el diagnóstico de los datos presión-gasto con un énfasis en la caracterización del mecanismo de producción del yacimiento. Recientemente, Bondarxvii resumió los métodos modernos e históricos de análisis de la relación agua-aceite (RAA) y los datos de flujo fraccional de agua (f_w), desde la perspectiva de caracterización del mecanismo(s) de empuje del yacimiento y la estimación de reservas.

A continuación se presenta un resumen breve del desarrollo de las curvas tipo de declinación de la producción.

Este desarrollo comienza en 1944 cuando Arps[iv] publicó sus curvas de declinación (hiperbólica, exponencial y armónica) para el análisis de los datos de producción. Debido a la simplicidad y consistencia de este acercamiento empírico, las ecuaciones de Arps mantienen su vigencia en la industria petrolera para el análisis e interpretación de los datos de producción.

El Análisis de Declinación de Arps, también conocido como Análisis Tradicional de Declinación, no requiere ningún parámetro del yacimiento ni del pozo, solamente debe aplicarse cuando el flujo alcanzó el límite del yacimiento y se tienen condiciones de operación constantes.

Arps presentó su método de análisis usando ecuaciones matemáticas que no tienen bases físicas, con excepción de la declinación exponencial, que muestra una tendencia de declinación. La función introducida por Arps está caracterizada por tres parámetros:

- Gasto inicial, q_i
- Declinación inicial, D_i
- Exponente de declinación, *b.*

Las relaciones de Arps para un pozo en producción son:

- Declinación Exponencial: (*b=0)*

$$q(t) = q_i * e^{-D_i t} \quad \text{.....} \quad (3.1)$$

- Declinación Armónica: (*b=1*)

$$q(t) = \frac{q_i}{1 + D_i t} \quad \text{.....} \quad (3.2)$$

- Declinación Hiperbólica: (*0<b<1*)

$$q(t) = \frac{q_i}{(1 + bD_i t)^{\frac{1}{b}}} \quad \text{.....} \quad (3.3)$$

Las relaciones empíricas de Arps se ilustran en la Figura 3.1, en donde se muestra en una gráfica las tres curvas de declinación para valores específicos de q_i y D_i, cada una de las curvas corresponde al comportamiento del gasto en el tiempo, para diferentes valores del exponente de declinación (*b*) y donde *D* es la declinación real del gasto.

En la Figura 3.2 se presentan en una gráfica doble logarítmica las curvas tipo, en donde es posible observar que a la declinación exponencial se le asocia un pronóstico de producción más conservador.

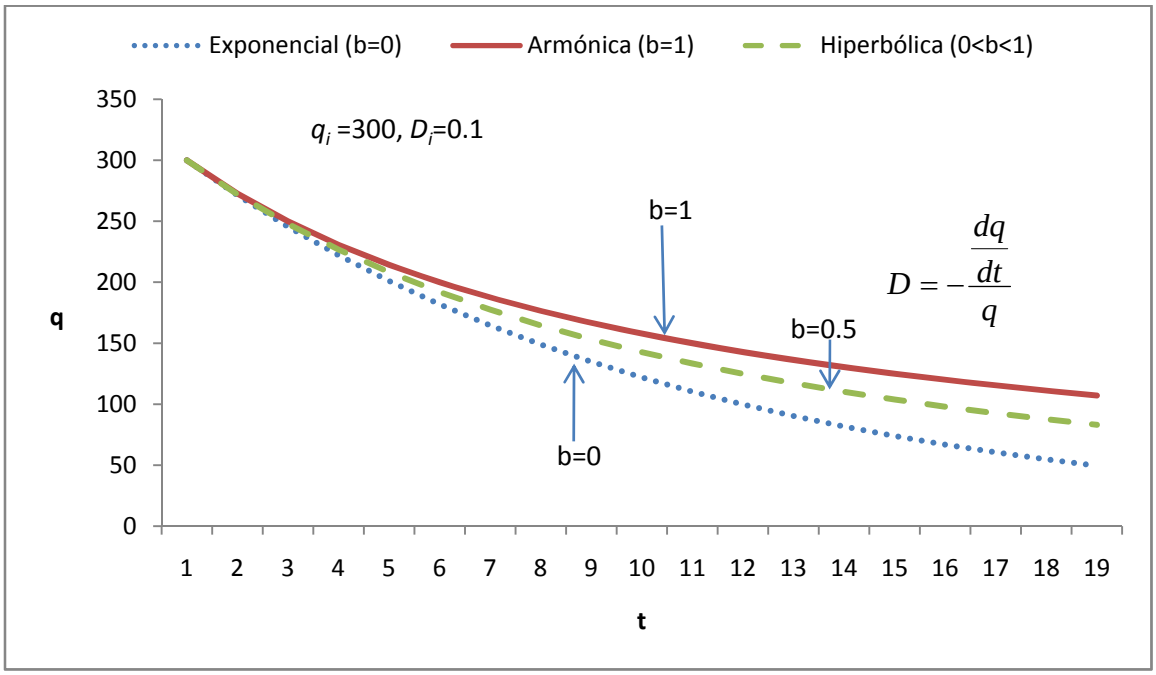


Figura 3.1. Curvas de declinación de Arps

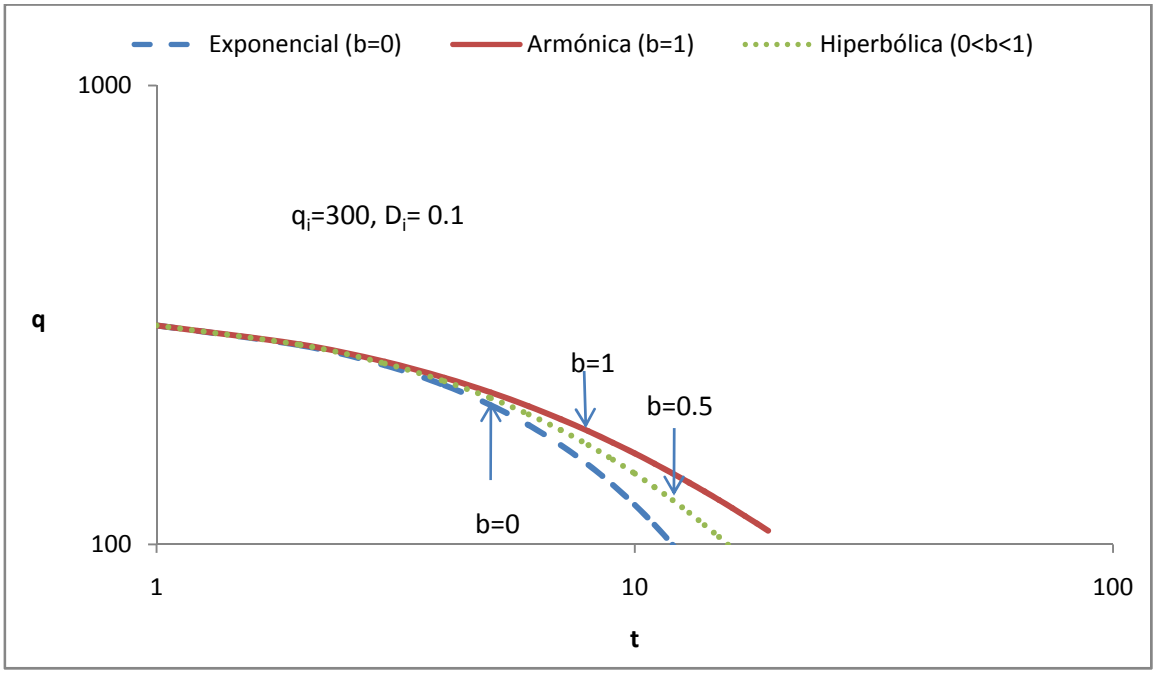


Figura 3.2. Curvas tipo empíricas de Arps

La ecuación de declinación hiperbólica es la más ampliamente usada, ya que las declinaciones exponencial y armónica constituyen casos especiales de la declinación hiperbólica, donde la razón de pérdida con respecto al tiempo adquieren valores de cero y uno, respectivamente.

Arps introdujo métodos de extrapolación de datos de producción contra tiempo, lo que permitió estimar reservas originales usando curvas de declinación exponenciales e hiperbólicas.

Nindxviii, mediante sus técnicas de graficación, mantuvo la popularidad de las relaciones de Arps. Es importante hacer evidente que tanto los estudios de Arps como el de Nind son empíricos, y que estos análisis de curvas de declinación no se relacionan con un modelo de yacimiento.

En 1980 Fetkovichxix introdujo el desarrollo más significativo de las curvas tipo de declinación; trató de relacionarlas a un modelo de yacimiento; utilizando una solución analítica unificada (declinación exponencial) para un pozo productor a presión de fondo constante, durante condiciones de flujo que ha alcanzado fronteras.

Fetkovich[xix] graficó las soluciones analíticas de declinación transitoria simultáneamente con las curvas de Arps, suponiendo que pueden usarse en un yacimiento no ideal (cambios de movilidad, yacimientos heterogéneos, con fracturas y capas múltiples). El resultado final de estas curvas es lo que se conoce como Curvas Tipo de Fetkovich, que proveen un análisis de los datos de producción durante condiciones de flujo tanto transitorio como cuando se sienten los efectos de la frontera; es decir, en el período donde un pozo tiene un comportamiento transitorio y su etapa posterior de declinación normal.

La metodología de Fetkovich analiza pozos de aceite produciendo a presión constante. Fetckovich combinó soluciones analíticas transitorias a tiempos cortos con ecuaciones de Arps para tiempos largos, que describen el flujo dominado por fronteras cerradas. Tanto el método de Fetkovich como la ecuación de Arps, calculan la recuperación final esperada.

El ajuste por curvas tipo es esencialmente una técnica gráfica para el ajuste visual de datos de producción, usando curvas pre-graficadas sobre papel doble logarítmico. Fetkovich usó las curvas de declinación de Arps junto con curvas tipo para flujo simétrico radial transitorio de líquidos ligeramente compresibles, a presiones de fondo constantes. Fetkovich[xix] relacionó los parámetros de declinación de Arps con algunos parámetros de ingeniería de yacimientos. Recomendó el uso de las curvas de declinación armónicas para pozos de gas. Fetkovichxx presentó una serie de ejemplos de campos petroleros empleando sus curvas de declinación; la conclusión más sobresaliente de este trabajo fue que las curvas de Arps no son válidas para el análisis de datos durante el período transitorio de la producción.

Fetckovich mostró que la familia de curvas empíricas de Arps puede combinarse con la solución de flujo de líquidos ligeramente compresibles, para obtener una familia de curvas que pueden usarse para predecir el comportamiento futuro y estimar el volumen poroso del yacimiento.

Fetkovich empleó ecuaciones de flujo analíticas para generar curvas tipo para flujo transitorio. Además, introdujo las variables adimensionales para el gasto y tiempo:

- Gasto $q_{Dd} = \frac{q(t)}{q_i}$; ..... (3.4)
- Tiempo $t_{Dd} = D_i t$ . ..... (3.5)

El autor obtuvo las relaciones siguientes (figuras 3.3 y 3.4):

- Exponencial: $q_{Dd} = e^{-t_{Dd}}$; ..... (3.6)
- Hiperbólica: $q_{Dd} = \frac{1}{(q + bt_{Dd})^{1/b}}$; ..... (3.7)
- Armónica: $q_{Dd} = \frac{1}{1 + t_{Dd}}$ . ..... (3.8)

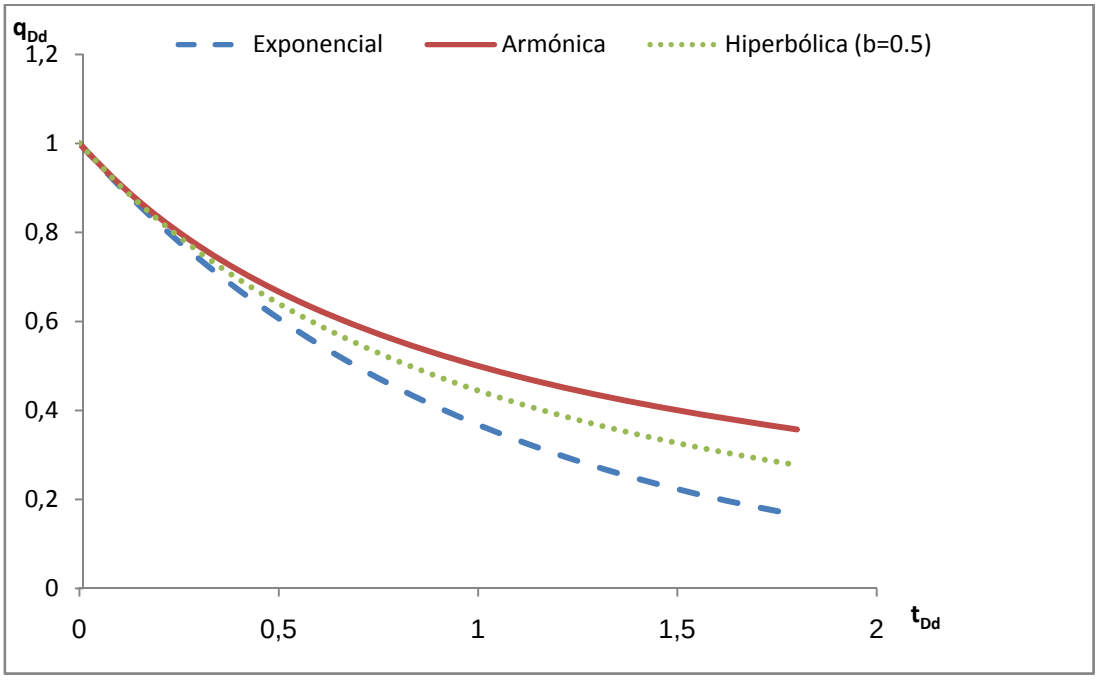


Figura 3.3 Declinación adimensional de Arps

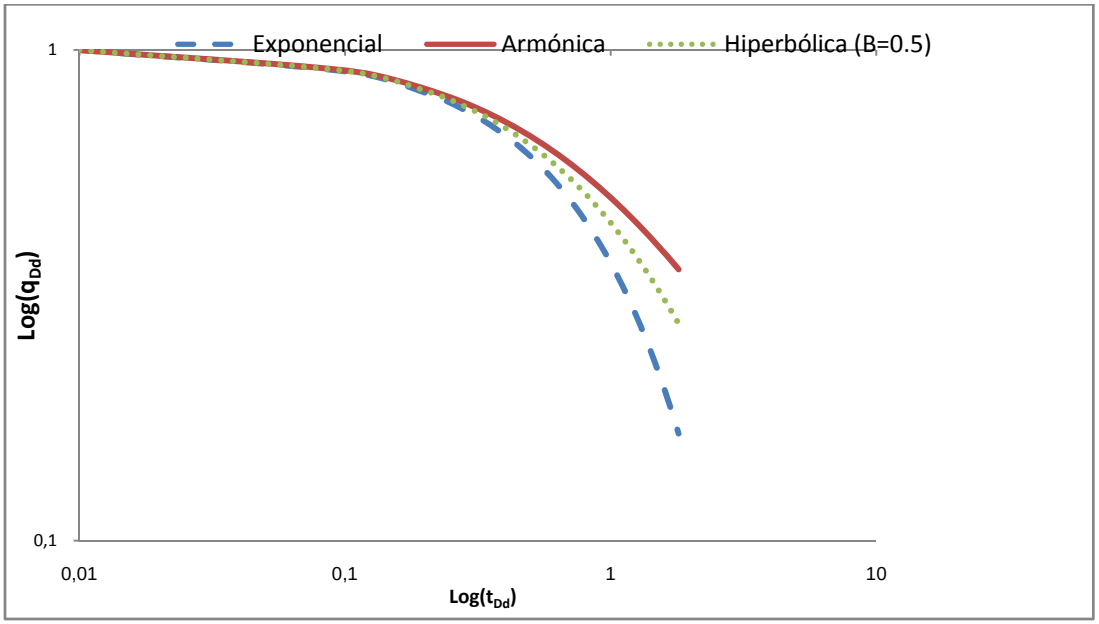


Figura 3.4.Declinación adimensional de Arps

Las curvas tipo de Fetkovich cubren la vida total de la producción de un pozo. Fetkovich dio un significado teórico al análisis de curvas de declinación.

Las curvas de Fetkovich son importantes en la ingeniería de yacimientos. Sin embargo, tienen limitaciones; la más común de ellas se presenta cuando los datos de producción muestran variaciones significativas de presión, ya sea por cierres o restricciones impuestas por condiciones operativas. Es por ello que varios estudios posteriores han tratado de superar este tipo de limitaciones y extender la aplicación de estas curvas. En 1986 Blasingame y Leexxi presentaron una propuesta para obtener el área de drene a partir de datos de producción variable.

En 1987 Fetkovich[xx] presentó una serie de ejemplos de campos empleando sus curvas de declinación.

Anderson y colaboradores[i], propusieron una serie de lineamientos para el análisis de datos de producción, con la expectativa de que al aplicarse en un conjunto de datos determinado, un analista pudiera ser capaz de:

- Garantizar la viabilidad de los datos, para lo cual se debe determinar cuando un conjunto de datos puede o no analizarse partiendo de la disponibilidad de la información necesaria:
    - Datos históricos de producción (gastos y presiones)

- Datos del yacimiento y de los fluidos (para un análisis cuantitativo)
- Registros de pozos (historia de terminaciones/estimulaciones)

- Verificar correlaciones de datos, el cual siempre es un paso intermedio entre la adquisición y el análisis, en el se debe efectuar una revisión final antes que los datos se procesen y preparen para el análisis. Las tareas recomendadas incluyen las siguientes:
  - Verificar correlación de datos (gráfica de p_{wf} o p_{tp} vs gasto). Esta es una validación simple, en donde datos que no tengan una correlación probablemente no proveerán ningún valor de diagnóstico
  - Gráficas gasto-tiempo y presión-tiempo, las cuales podrán mostrar características o eventos que deberán ser filtrados o descartados
- Realizar un diagnóstico preliminar, tarea que considera los aspectos de diagnóstico siguientes:
  - Identificar regímenes de flujo (modelo del yacimiento)
  - Filtrado de datos para claridad (eliminación de datos malos)
  - Revisión/edición de datos
- Realizar un análisis basado en el modelo, que de alguna forma es la parte más fácil del análisis de los datos de producción, similar al análisis de pruebas de pozos; comparación/ajuste con modelos de pozo/yacimiento, refinamientos, y si se requiere, predicción de la producción.

Los lineamientos propuestos por Anderson y colaboradores[i] para el análisis de producción se resumen en las los siguientes:

- Revisión de la historia de producción con fines de consistencia.
- Revisión de la historia del pozo, particularmente terminaciones, reparaciones, estimulaciones.
- Reunir datos del yacimiento (datos de registros geofísicos, de estudios petrofísicos en núcleos y datos de fluidos, PVT)
- Realizar análisis de diagnóstico de los datos de producción
  - Revisión de la historia de datos y validar su correlación de datos (datos de gasto y presión)
  - Realizar un análisis simplificado de los datos de producción

- Establecer un modelo de yacimiento usando gráficas de diagnóstico

- Ejecutar un análisis basado en el modelo/predicción de los datos de producción.

El concepto de gráfica de diagnóstico implica que una cierta característica o comportamiento emergerá de un perfil de datos determinado. Anderson y colaboradores[i] propusieron las gráficas de diagnóstico para el análisis de producción incluidas en la

Tabla **3.1**.

Tabla 3.1. Resumen de las diferentes gráficas de diagnóstico para el análisis de datos de producción

Gráfica de diagnóstico	Variables para aceite	Variables para gas	Valor en la práctica
Historia y correlación de datosiv,xvi,xvii			
	log*(q)* y p_{wf} vs *t*	q_g	Bueno
	log*(q)* y log(N_p) vs Log*(t)*	q_g,G_p	Bueno
	p_{wf} vs *q*	q_g	Moderado
	q y p_{wf} vs N_p	q_g,G_p	Moderado
Diagnóstico de yacimientoix,x,xi			
	log(Δp) vs log(N_p/q)	$(\Delta m(p), q_g, G_p$	Muy bueno
	log(q/Δp*)* vs *log*(N_p/q)	$(\Delta m(p), q_g, G_p$	Muy bueno
Diagnostico auxiliarv,xxii,xxiii			
	log(q/Δp*)* vs log$(N_p/\Delta p)$	$(\Delta m(p), q_g, G_p$	Bueno
	log*(1/q)* vs. log(N_p/q)	q_g, G_p	Muy bueno
	*$(p_{wf})_{medida}$ y $(p_{wf})_{decon}$ vs *t*		Bueno

*Requiere transformaciones de pseudo-presión y pseudo-tiempo para la deconvolución.

.

Estas gráficas de diagnóstico deberán resaltar aquello que sea erróneo en los datos de producción; es decir, identificar las causas del mal comportamiento de la gráfica (aunque este es un proceso más cualitativo que analítico, correspondiente a la naturaleza de la gráfica de diagnóstico) y verificar la correlación o carencia de un conjunto de datos de gasto y presión.

Los autores también nos muestran un ejemplo (Tabla 3.2) de los retos y problemas que frecuentemente se presentan en el análisis e interpretación de los datos de producción, que deben tomarse en cuenta ya sea para reconocerlos o para corregirlos. En el desarrollo de este trabajo es importante conocer el grado de influencia que pueden tener en nuestro análisis, para realizar las correcciones pertinentes.

Tabla 3.2. Retos y problemas comunes en el análisis de datos de producción.

Problema		Severidad/influencia
Presión	No existen mediciones	Alta
	Estimación de presión inicial incorrecta	Alta
	Modelo de conversión $p_{tp} \rightarrow p_{wf}$ deficiente	Moderada
	Almacenamiento de líquidos en el pozo: efecto sobre la conversión $p_{tp} \rightarrow p_{wf}$	Moderada
	Localización incorrecta de las mediciones de presión	Muy alta
Gasto	Ausencia de gastos	Moderada
	Almacenamiento de líquidos en el pozo: efecto de flujo de gas	Moderada
Terminaciones	Cambios en la zona aledaña: perforaciones nuevas o viejas	Muy alta
	Cambios en la zonas tubular del pozo	Alta
	Cambios en equipo superficial	Moderada/Alta
	Estimulación: fracturamiento hidráulico	Alta
	Estimulación: acidificación, etc.	Moderada
General	Propiedades del yacimiento	Moderada
	Propiedades del aceite y/o gas	Moderada
	Mala sincronización de la relación gasto-presión-tiempo	Moderada/Alta
	Mala correlación de la relación tiempo-presión-gasto	Muy alta

Un trabajo reciente de Ilk, Mattar y Blasingamexxiv coincide en que las gráficas de diagnóstico específicas, actualmente en uso para el análisis de datos de producción, incluyen:

- Gráfica de historia de producción.
- Gráfica de correlación presión-gasto.
- Gráfica Log-Log o del Índice Normalizado de Productividad: log ($\Delta p/q$) vs log (*función de tiempo de balance de materia*). La referencia más relevante para este enfoque es Agarwal, y colaboradores[xi], mencionada anteriormente.
- Gráfica Blasingame o Curva Tipo Avanzada de Declinación: log ($q/\Delta p$) vs. log (*t* de balance de materia).[ix,x]

Puede observarse la coincidencia con las gráficas de diagnóstico mencionadas en la tabla 3.2.

Podemos concluir que las gráficas de diagnóstico son indispensables en el análisis de los datos de producción, así como que la historia de producción es un elemento esencial del proceso de análisis e interpretación y siempre debe tomarse en cuenta, sobre todo en terminaciones, estimulaciones y/o reparaciones mayores. Finalmente, cualquier desajuste en una gráfica de diagnóstico particular puede no ser un signo de falla eminente en el análisis o interpretación. Sin embargo, cualquier desajuste deberá investigarse a fondo.

Es importante supervisar la adquisición de los datos, particularmente los datos de presión; realizar revisiones periódicas de datos, incluyendo la posibilidad de pruebas de pozos para evaluar las condiciones de producción presentes del pozo; y, realizar esfuerzos continuos en el desarrollo de gráficas de diagnóstico de datos, particularmente gráficas que permitan asegurar la correlación de datos de presión y gasto.

,

CAPÍTULO 4. Métodos de Minería de Datos

Este capítulo presenta una descripción general de algunos de los métodos de minería de datos que existen, los cuales se utilizarán en el desarrollo de este trabajo. Generalmente estos métodos se activan o conducen por datos, lo que significa que los algoritmos se usan para encontrar estructuras en los datos, sin retroalimentación del ingeniero. Cuando se trabaja con grandes cantidades de datos, ésta es una característica importante. Estos métodos pueden usarse solos o en conjunto, y ser útiles en tareas de modelado, predicción, clasificación y optimización.

En años recientes, han aparecido diferentes herramientas de minería de datos. Estas consisten básicamente en programas de cómputo que permiten transformar y manipular conjuntos de datos. Aunque ésta es su función, el éxito de estas herramientas requiere de experiencia y conocimiento. Además, se necesitan herramientas de apoyo tales como bases y almacenes de datos.

Algunas de las herramientas de minería de datos más comúnmente usadas en la ingeniería petrolera son las redes neuronales, los algoritmos genéticos, los algoritmos de lógica difusa y los mapas auto - organizados. En este trabajo se describirán las redes neuronales y los algoritmos de lógica difusa, ya que son las herramientas que se aplicarán para desarrollar el método de validación de datos.

En la primera sección de este capítulo se describirán las redes neuronales artificiales. Después se continuará con la descripción de los algoritmos basados en lógica difusa, y finalmente se discutirán los algoritmos genéticos.

4.1. Teoría de Redes Neuronales

Una red neuronal artificial es un elemento capaz de procesar una gran cantidad de información en forma paralela y distribuida, inspirada en la redes neuronales biológicas, las cuales pueden almacenar conocimiento y tenerlo disponible para su usoi. Éstas tienen algunas semejanzas con el funcionamiento del cerebro humano, como son:

1. El conocimiento se adquiere a través del proceso de aprendizaje.

2. La conectividad entre neuronas se determina por los pesos sinópticos, lo cuales se utilizan para almacenar el conocimiento.

El cerebro tiene algunas características deseables para cualquier sistema de procesamiento digitalii:

1. Es robusto y tolerante a fallas; diariamente mueren neuronas sin afectar su desempeño.
2. Es flexible; se ajusta a nuevos ambientes por medio de un proceso de aprendizaje y no hay que reprogramarlo.
3. Puede manejar información difusa, con ruido o inconsistencias.
4. Es altamente paralelo.
5. Es pequeño, compacto y consume poca energía.

Basados en la eficiencia de los procesos llevados a cabo por el cerebro e inspirados en su funcionamiento, varios investigadores han desarrollado desde hace mas de 30 años la teoría de las Redes Neuronales Artificiales (RNA), las cuales emulan el comportamiento de las redes neuronales biológicas, que se han utilizado para aprender estrategias de solución basadas en ejemplos de comportamiento típico de patrones; estos sistemas no requieren que la tarea a ejecutar se programe, debido a que aprenden de la experiencia.

La función del proceso de aprendizaje es modificar los pesos sinópticos de las redes neuronales con el fin de minimizar una función objetivo. La modificación de los pesos sinópticos es el método tradicional para el diseño e implementación de las redes neuronales.

Brevemente se describe una red neuronal biológica. Una neurona típica contiene un cuerpo (donde está localizado el núcleo), dendritas, y un axón. La información entra al cuerpo de la célula en forma de pulsos electroquímicos entrenados (señales), a través de las dendritas. Dependiendo de la naturaleza de la entrada, la neurona se activa en una forma excitatoria o inhibitoria y entrega una salida, que viaja a través del axón y conecta a las otras neuronas, donde ésta se convierte en la entrada de la neurona receptora. El punto de unión entre las dos neuronas es una ruta neural, donde la terminación del axón de una neurona está muy cerca del cuerpo de la otra neurona, o

con las dendritas. Estas conexiones se llaman sinapsis. Las señales viajando desde la primera neurona inician un entrenamiento del punto electroquímico (señales) en la segunda neurona.

Las redes neuronales artificiales son sistemas de proceso de información, que representan una aproximación burda y simplificada para la simulación del proceso biológico, intentando obtener características de desempeño similares.

Zangl**¡Error! Marcador no definido.** menciona que las habilidades para aprender, asociar, y ser tolerante a errores, son características importantes de las redes neuronales. Las redes neuronales pueden reconocer datos incompletos o erróneos, lo cual es benéfico para la predicción, diagnóstico o control de procesos. Las redes neuronales deben aplicarse cuando el ingeniero requiere generar un modelo de un problema con una relación no lineal entre la entrada y la salida.

Las redes neuronales son herramientas poderosas debido a que pueden aprender y adaptarse, y toleran eventos incompletos o anómalos.

Una red neuronal artificial es una colección de neuronas, las cuales son las unidades básicas de procesamiento y normalmente están organizadas en capas. La información se propaga a través de la red neuronal, capa por capa, siempre en la misma dirección. Entre la capa de entrada y la de salida hay otros niveles intermedios de capas de neuronas, llamadas capas ocultas. La Figura 4.1 muestra la arquitectura de una red neuronal.

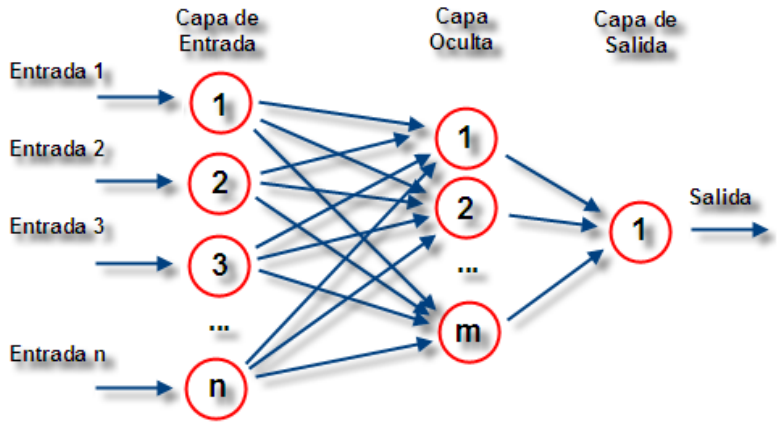


Figura 4.1. Arquitectura de una red neuronal con una neurona en la capa de salida

Una red neuronal artificial normalmente consiste de una capa de entrada, de una o más capas ocultas y una capa de salida. El número de neuronas en la capa de entrada, corresponde al número de parámetros que están siendo presentados a la red como entrada. Lo mismo es válido para la capa de salida. El análisis de la red neuronal no está limitado a una salida única y las redes neuronales pueden entrenarse para construir modelos neuronales con salidas múltiples. Las neuronas en la(s) capa(s) intermedias ocultas son responsables de la característica de extracción, proporcionan una capacidad de incrementar las dimensiones y resuelven tareas, tales como la clasificación y el reconocimiento de patrones.

La neurona es la unidad fundamental para la operación de la red neuronal. Una red neuronal está compuesta por:

1. Un conjunto de uniones o conexiones sinápticas, con cada elemento caracterizado por su propio peso.
2. Un sumador, el cual incluye los componentes de la señal de entrada multiplicado por su peso respectivo.
3. Una función de activación no-lineal que transforma la salida del sumador en la entrada de la neurona siguiente.

Al esquema de la red neuronal también se aplica un umbral para reducir la entrada a la función de activación. En términos matemáticos, la i-ésima neurona se puede describir como:

$$u_i = \sum_{j=1}^{n} w_{ij} x_j \qquad \text{..... (4.1)}$$

Donde:

x_j : j-ésimo componente de la entrada
w_{ij} : peso de la conexión entre el j-ésimo componente de la entrada y la i-ésima neurona
u_i : salida del sumador

En la Figura 4.2 se muestra el modelo no lineal de una i-ésima neurona, en donde:

ρ_i : umbral,
$\varphi(.)$: función de activación no lineal,
y_i : salida de la i-ésima neurona.

La función de activación no-lineal denotada por $\varphi(.)$, recibe como entrada u_i y genera el elemento de la salida y_i como se describe en la ecuación siguiente:

$$y_i = \varphi(u_i)$$

Una clasificación de este tipo de función es:

1. Diferenciable y no diferenciable
2. Tipo pulso y tipo escalón
3. Positiva y de promedio cero

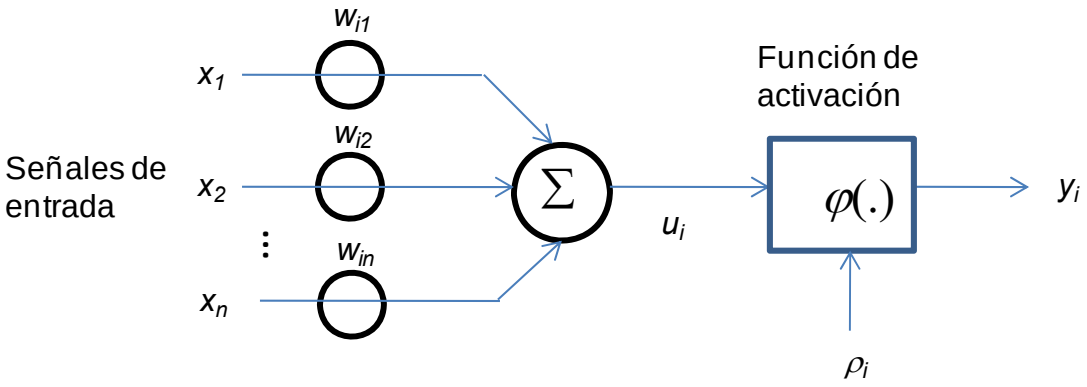
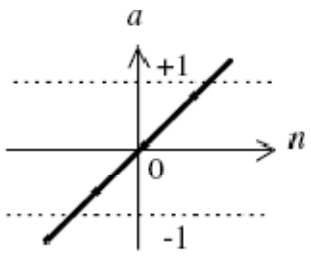


Figura 4.2. Modelo no lineal de una neurona

Las funciones principales de activación empleadas en redes neuronales son:

1. Lineal (purelin).

   La salida de una función de activación lineal es igual a su entrada,

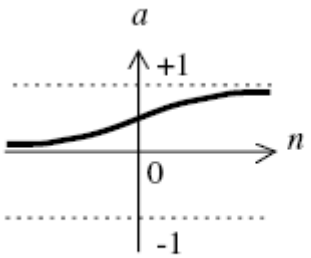


$a = purelin(n)$

$a = n$

2. Sigmoidal (logsig) o logística.

La función sigmoidal toma la entrada, que puede tener un valor entre -∞ y +∞, y entregar una salida entre 0 y 1. Esta función de activación se usa en los algoritmos de aprendizaje con retro-propagación del error, en parte porque es diferenciable.



$a = logsig(n)$

$$a = \frac{1}{1 + e^{n}}$$

Uno de los aspectos más importantes de las RNA es el almacenamiento de información. Cada unas de las conexiones entre neuronas está equipada con un peso individual, como se puede observar en la Figura 4.2, que modifica la señal en esa conexión. El peso sirve como un factor por el cual la salida de la neurona precedente se multiplica antes de convertirse en una entrada. Esto significa que la información se almacena y distribuye dentro de una RNA, y la alteración pequeña de algunos de los pesos tendrá un efecto reducido al retomar la información aprendida.

Las RNA es un método de reconocimiento de patrones que requiere de un conjunto de datos representativo. Las RNA descubren la influencia de cada una de las entradas en las salidas de un proceso iterativo.

La fase en la que la RNA utiliza la muestra de datos representativa de un problema para reconocer patrones, se conoce como fase de entrenamiento. Durante la fase de entrenamiento, los pesos de la RNA se ajustan. Dependiendo del tipo de red neuronal y del problema que se pretende resolver, un método supervisado o no puede utilizarse para ajustar los pesos. Sin embargo, en ambos casos, el entrenamiento inicia con una primera asignación de pesos; la entrada se propaga a través de la RNA y en consecuencia todas sus neuronas cambian su actividad.

Un uso común para las redes neuronales es la clasificación de datos. Esta tarea se inicia cuando un conjunto representativo de ejemplos describiendo un problema se introduce en el sistema. La red de neuronas empieza a extrapolar el mapeo entre los datos de entrada y de salida. Después de entrenarse, el sistema puede usarse para reconocer datos que son similares a algunos de los ejemplos empleados durante la fase de entrenamiento. La red neuronal puede reconocer datos incompletos o contaminados. Esta es una característica importante, usada frecuentemente para fines de predicción, diagnóstico o control.

Otra tarea que puede manejarse efectivamente por las redes neuronales es la estimación de datos faltantes. Este método puede usarse para datos que obedecen ciertas relaciones, especialmente de tipo no lineal. Esto permite determinar los valores faltantes a pesar de la dependencia no lineal de las variables.

Existen tres clases de arquitecturas de redes neuronales artificiales: capa simple, multicapa y recurrente. Una red de una capa se ilustra en la Figura 4.1, en donde solo existe una capa oculta de *m* neuronas, en la que cada una de las *n* entradas se conectan a cada una de las neuronas. En este caso la matriz de pesos tiene *m* renglones. La capa incluye además de la matriz de pesos, los sumadores, el vector de umbrales, la función de activación y el vector de salida. En el caso de la red multicapa, cada capa oculta tendrá su propia matriz de pesos, su vector de umbrales, un vector de entradas y un vector de salidas. Las redes recurrentes son un poco diferentes, ya que

contienen una realimentación hacia atrás o retroalimentación, es decir, algunas de sus salidas son conectadas a sus entradas.

4.2. Algoritmo de Retro-propagación

La red neuronal artificial más común que usa entrenamiento supervisado es llamada red multicapa con retro-propagación del aprendizaje. Esta red es la más empleada en clasificación, además de ser la más simple.

El nombre retro-propagación se deriva del método por el cual se corrigen los pesos. Durante la fase de aprendizaje, los patrones de entrada se presentan a la red de acuerdo a una secuencia determinada. Cada uno de los patrones de entrenamiento se propaga hacia adelante, capa por capa, hasta que se calcula una salida. La salida calculada se compara con la salida objetivo o deseada y con base en esta comparación se determina un error. Los errores se usan como entradas para retroalimentar las conexiones, ajustando los pesos sinápticos hacia atrás, capa por capa.

El proceso de entrenamiento por retro-propagación del error de una red involucra tres fases: la propagación hacia adelante de los patrones de entrenamiento de entrada (*feedforward*), el cálculo y la propagación hacia atrás del error asociado, y el ajuste de los pesos. Después del entrenamiento, la aplicación de la red incluye solamente los cálculos de la primera fase (*feedforward*). Aún si el entrenamiento es bajo, una red entrenada puede producir su salida muy rápidamente. Con el propósito de mejorar la velocidad del proceso de entrenamiento, se han desarrollado numerosas variaciones de retro-propagación.

Se puede establecer en forma resumida, que una red con un sistema de entrenamiento mediante retro-propagación consiste en:

- Empezar con unos pesos sinápticos cualquiera (generalmente elegidos al azar).
- Introducir algunos datos de entrada (en la capa de entradas) elegidos al azar, entre los datos de entrada que se van a usar para el entrenamiento.
- Dejar que la red genere un vector de datos de salida (propagación hacia delante).
- Comparar la salida generada por la red con la salida deseada.

- La diferencia obtenida entre la salida generada y la deseada (denominada error), se usa para ajustar los pesos sinápticos de las neuronas de la capa de salidas.
- El error se propaga hacia atrás (retro-propagación), hacia la capa de neuronas anterior, y se usa para ajustar los pesos sinápticos en esta capa.
- Se continúa propagando el error hacia atrás y ajustando los pesos, hasta que se alcanza la capa de entrada.

Este proceso se repetirá con los diferentes datos de entrenamiento hasta alcanzar un criterio de convergencia.

Como se ha mencionado, el modelo de retro-propagación es un algoritmo que consiste en minimizar un error (frecuentemente cuadrático), por medio de un gradiente descendiente. Algunos de los parámetros que pueden permitir la optimización del proceso de aprendizaje son el momento, el decaimiento del peso y la razón de entrenamiento, los cuales se explican a continuación.

Momento. Es un empuje extra al proceso de aprendizaje que tiene dos propósitos. Primero, puede acelerar el proceso de aprendizaje, y segundo, tiene el potencial para impulsar la solución de los mínimos locales, que usualmente existen en el espacio de búsqueda, lo que ayuda a que las soluciones converjan rápidamente.

Razón de aprendizaje. Es un indicador de que tan rápido se requiere que la red aprenda; generalmente tiene un valor entre 0 y 1. Un valor de razón de aprendizaje alto puede provocar que la red pierda el mínimo global en el espacio de búsqueda y podría causar problemas en la convergencia durante el entrenamiento. Un valor de razón de aprendizaje bajo puede prolongar el proceso de aprendizaje, y reducirlo a un rastreo.

Decaimiento del peso. Agrega un término de penalización a la función de error. La penalización normal es la suma de los cuadrados de los pesos como una constante de decaimiento.

4.3. Lógica Difusa

Una de las disciplinas matemáticas con mayor número de seguidores actualmente es la llamada lógica difusa o borrosa, la cual utiliza expresiones que no son ni totalmente ciertas ni completamente falsas; es decir, es la lógica aplicada a conceptos que pueden

tomar un valor cualquiera de veracidad dentro de un conjunto de valores que oscilan entre dos extremos, la verdad absoluta y la falsedad totaliii. Conviene recalcar que lo que es difuso, borroso, impreciso o vago no es la lógica en sí, sino el objeto que estudia: expresa la falta de definición del concepto al que se aplica. La lógica difusa permite tratar información imprecisa como “temperatura media” o “temperatura baja”, en términos de conjuntos que se combinan en reglas para definir acciones: *si la temperatura es alta entonces enfriar mucho*. De esta manera, los sistemas de control basados en lógica difusa combinan variables de entrada, definidas en términos de conjuntos difusos, por medio de grupos de reglas que producen uno o varios valores de salida.

El primer trabajoiv relacionado con el tema de la “vaguedad” data de la primera década del siglo 20, cuando el filósofo americano Piercev notó que “eliminar la vaguedad de la lógica es como eliminar la fricción en mecánica”. A principios de 1920, el matemático y lógico polaco Lukasiewiczvi desarrolló la lógica de tres valores y habló de lógica de valores múltiples. En 1937, el filósofo cuántico Blackvii publicó un artículo sobre conjuntos “vagos”. Estos científicos construyeron las bases sobre las cuales se desarrolló la nueva lógica difusa.

Zadehviii, conocido como el padre de la lógica difusa, publicó su artículo de arranque “Fuzzy Sets” en 1965. Desarrolló muchos conceptos importantes, incluyendo el de valores de pertenencia, y proporcionó una base amplia para aplicar la teoría a problemas de ingeniería y científicos. Esta base incluyó las operaciones clásicas para conjuntos difusos, las cuales comprenden todas las herramientas necesarias para aplicar la teoría de los conjuntos difusos a problemas del mundo real. Zadeh fue el primero en usar el término “fuzzy” (difuso), el cual provocó mucha oposición. A pesar de todos sus adversarios, la lógica difusa continuó creciendo y se ha convertido en una fuerza importante tras de muchos avances en sistemas inteligentes.

En esencia, la lógica difusa proporciona los medios para realizar cálculos con palabras. Usando lógica difusa, los expertos ya no estarán obligados a concentrar su conocimiento a un lenguaje que las computadoras puedan entender. La lógica difusa está soportada en la teoría de los conjuntos difusos.

4.4. Teoría de Conjuntos Difusos

Los conjuntos difusos son una manera de representar la incertidumbre. La incertidumbre usualmente es el resultado ya sea de la naturaleza aleatoria de los eventos, o de la imprecisión y ambigüedad de la información que tenemos acerca de los problemas cuando estamos tratando de resolverlos. En un proceso aleatorio, el resultado de un evento de entre varias posibilidades, es estrictamente el resultado de la oportunidad. Cuando la incertidumbre es un producto de la aleatoriedad de eventos, la teoría de la probabilidad es la herramienta apropiada a usarse. Las observaciones y medidas pueden emplearse para resolver incertidumbre aleatoria o estadísticas. Por ejemplo, una vez que la moneda se arroja, ya no existe incertidumbre aleatoria o estadística.

La mayoría de las incertidumbres, especialmente cuando se trabaja con sistemas complejos, resultan de la carencia de información. El tipo de incertidumbre que produce la complejidad de un sistema viene de la imprecisión, de la habilidad limitada para realizar mediciones adecuadas, de la carencia de conocimiento, o de la vaguedad (tal como la vaguedad inherente en el lenguaje natural). La teoría de conjuntos difusos es una herramienta maravillosa para el modelado de cierta clase de incertidumbre asociada con la vaguedad, imprecisión, y/o la carencia de información, correspondiente a un particular elemento del problema que tenemos. La lógica difusa consigue esta importante tarea a través de los conjuntos difusos. En conjuntos exactos, un objeto siempre pertenece o no a un conjunto determinado. En conjuntos difusos, cualquier cosa es asunto de grados. Así, un objeto pertenece a un conjunto con un cierto grado de certidumbre.

Clasificación difusa

Clasificar es el agrupamiento de objetos similares. En otras palabras clasificar es el proceso de agrupar elementos de un conjunto de datos, dentro de grupos (o clases) de acuerdo a criterios similares. Uno de los métodos conocidos de clasificación Fuzzy C-Means, o clasificación difusa, es una técnica usada para agrupar un conjunto de datos dentro de una clase o grupo, de tal forma que los elementos en la misma clase tienen

un alto grado de similitud, mientas que elementos perteneciendo a diferentes clase tienen un alto grado de diferencias.

Este método de clasificación está soportado por la lógica difusa, que como mencionamos en párrafos anteriores, es la lógica que utiliza expresiones que no son ni totalmente ciertas ni completamente falsas; es decir, es lógica aplicada a conceptos que pueden tomar un valor cualquiera de veracidad dentro de un conjunto de valores que oscilan entre dos extremos, la verdad absoluta y la falsedad total[viii].

Entropía. Por definición, es la medida de la carencia de orden en un sistema. Sin embargo, en Clasificación Difusa, la entropía mide el grado de carencia de similitud entre elementos. La entropía de un elemento o dato se entiende como la relación entre el valor de pertenencia difusa mínimo y el valor de pertenencia difusa máximo del elemento o dato.

4.5. Algoritmo de Clasificación Fuzzy C-Means

Como se mencionó, Fuzzy C-Means (FCM) es una técnica de clasificación de datos en la que cada uno de los puntos pertenece a una clase o grupo en algún grado, y éste que es conocido como el grado de pertenencia. En este método un elemento puede pertenecer a una o más clases. Esta técnica fue introducida inicialmente en 1981 por Jim Bezdek, como una mejora a los métodos de clasificación existentes. Este método muestra como agrupar puntos de datos de un espacio multidimensional, dentro de un número específico de clases diferentes. FCM es frecuentemente usada en reconocimiento de patrones.

Este método está basado en la minimización de la función objetivo siguiente:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad ,1 \leq m < \infty, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.2)$$

donde el parámetro de difusividad, m, el es un número mayor a 1, usado para ajustar el efecto de los valores de pertenencia; u_{ij} es el grado de pertenencia de x_i en la clase j, x_i es el i-ésimo elemento del conjunto de d dimensiones, c_j es el centro de d dimensiones de la clase, y $\|*\|$ es cualquier norma que exprese la semejanza entre el dato y el centro de la clase.

La partición difusa se realiza a través de un proceso iterativo de optimización de la función objetivo con la actualización de u_{ij} y el centro de la clase c_j:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right]^{\frac{2}{m-1}}} \quad , \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.3)$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad . \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.4)$$

El proceso iterativo se detendrá cuando $max_{ij} = \{|u_{ij}^{k+1} - u_{ij}^{k}|\} < \varepsilon$, donde ε es un criterio entre 0 y 1 y k es el contador de iteraciones. Este procedimiento converge a un valor mínimo, $J_{m.}$

El algoritmo consta de las etapas siguientes:

1. Inicializar una matriz $U = [u_{ij}]$, $U(0)$.
2. En el k-ésimo paso calcular los vectores centro $C^{(k)} = [c_j]$ con $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad , \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.5)$$

3. Actualizar $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right]^{\frac{2}{m-1}}} \quad . \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.6)$$

4. Si $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$ concluye el proceso; si no se cumple se regresa al punto 2.

Como se mencionó previamente, los datos se asocian a cada una de las clases por medio de una función de pertenencia, que representa el comportamiento difuso de este algoritmo. Para realizar lo anterior, simplemente se tiene que construir una matriz

apropiada denominada U, cuyos componentes son números entre 0 y 1, y representan el grado de pertenencia entre el dato y los centros de las clases.

Este tipo de algoritmos han sido aplicados en la industria petrolera en los estudios de petrofísica, recuperación mejorada, así como en perforación de pozos. En este trabajo de tesis se empleará un sistema híbrido que incluye el algoritmo Fuzzy C-Means y una red neuronal artificial.

CAPÍTULO 5. Técnicas de minería de datos para validación del dato de producción

Actualmente, el problema de disponer de datos de calidad requiere más atención que la que normalmente se le brinda. Para entender lo anterior, se definen los datos *contaminados* o *corruptos* como aquellos que se interpretan o registran incorrectamente en las bases de datos. Es importante recordar que el éxito de la industria petrolera se apoya en la combinación de información confiable multidimensional y en el análisis de esta información con técnicas propiamente interdisciplinarias, y que los datos contaminados pueden conducir a la toma inadecuada de decisiones a nivel estratégico, lo que puede causar fallas operacionales.

La calidad de los datos se define como la medida de la coincidencia entre las vistas de datos presentadas por los sistemas de información y la misma información en el mundo real. Si la coincidencia no es sustancial para algún período de tiempo, entonces el sistema de información está mal construido.i

La idea esencial tras esta validación de datos es el viejo axioma en computación: “basura que entra-basura que sale”. La validación analítica se centra en determinar cuando los datos disponibles son basura, buenos o algo intermedio.

En este trabajo se propone una nueva metodología para la validar la calidad de los datos de producción. Dicha metodología se basa en la hipótesis siguiente: en un sistema con un comportamiento adecuado, la salida deberá ser capaz de contribuir a su propia predicción e identificación.

5.1. Sistema de clasificación del dato de producción

La metodología propuesta usará como herramientas las técnicas de clasificación o agrupamiento difuso, modelado de redes neuronales y un proceso iterativo para alcanzar una meta convergente. El resultado será la clasificación de los datos de producción en *bueno*, *contaminado* o *ligeramente contaminado*.

Para la aplicación de esta metodología se usarán los parámetros obtenidos durante el proceso de producción. Primero se empleará una técnica de agrupamiento difuso para

la clasificación de los datos. La salida del sistema, en este caso el gasto de aceite, se incluirá en el conjunto de datos. De esta manera, la salida participará indirectamente en el modelo de la red neuronal, y así la información generada por el análisis estará considerando el sistema completo (entradas y salidas). Cada uno de los elementos del conjunto de datos se clasificará en tres grupos o clases difusas, y se determinará el grado de pertenencia de cada uno de ellos, a cada una de las tres clases. Con esta información se calculará la entropía de cada elemento, con lo que se podrá identificar el nivel de caos de cada uno de ellos. La información del grupo o clase, conjuntamente con la entropía, se agregará a los parámetros del conjunto de datos y se entrenará la red neuronal. Con el fin de concluir cuando un dato será clasificado como bueno, contaminado o ligeramente contaminado, se desarrollarán tres curvas, las cuales se identificarán para cada uno de los puntos (valor discreto de la salida en *X*), la información de la clase o grupo predominante (curva 1), la entropía (curva 2), y la diferencia entre el valor real y el valor estimado por la red neuronal (curva e3). Así, la curva 1 tendrá siempre un valor discreto, ya sea 1, 2 ó 3 y se representará por medio de una función escalón. La curva 2 será una función continua que siempre tendrá un valor entre cero y uno. La curva 3 mostrará que tan cercana es la predicción de la red neuronal con respecto al valor real. La clave en este proceso será la presencia de datos de salida (gasto de aceite) en el proceso de agrupamiento difuso, ya que el gasto de aceite se usa durante el entrenamiento de la red neuronal a través de los valores del grado de pertenencia, la clase y la entropía. Esto significa que la salida del sistema está contribuyendo indirectamente en la entrada. A continuación se describe el algoritmo de validación de datos.

Algoritmo. Los pasos involucrados en la clasificación de datos son los siguientes (Figura 5.1):

1. Clasificar los elementos del conjunto de datos a validar, incluyendo el parámetro de salida del sistema.
2. Entrenar una red neuronal usando la información de clasificación generada en el paso 1 (clase predominante, grado de de pertenencia de esa clase dominante y entropía).
3. Iterar mediante un proceso que consiste en:

a. Seleccionar un elemento para el rango de la salida del sistema
b. Recorrer el rango de salida mientras se corre la red
c. Para cada uno de los valores de salida supuestos:
    i. Calcular el número de clase dominante
    ii. Calcular grado de pertenencia para la clase dominante
    iii. Calcular la entropía como la razón entre el grado de pertenencia máximo entre el grado de pertenencia mínimo
    iv. Correr la red neuronal con estos tres parámetros adicionales
    v. Calcular el error:

        Error = |Salida de la Red Neuronal – Salida Supuesta|
d. Generar tres curvas, en donde cada uno de sus puntos corresponde a un valor de salida supuesto.
    i. Curva 1 o curva de clase: Identifica la clase a la cual pertenece el punto.
    ii. Curva 2 o curva de entropía: Identifica la entropía del dato dentro de la clase a la que pertenece.
    iii. Curva 3 o curva de error: Identifica la diferencia entre el valor real y el valor que predice la red neuronal.
e. Desplegar las tres curvas en una gráfica.
f. Identificar para cada una de las clases, la posición donde el error se aproxime al valor de cero y la entropía sea mínima.
g. En la misma gráfica del punto e, representar como líneas verticales el valor de salida original y el valor de predicción de la red neuronal del elemento que se está evaluando.

4. Repetir el proceso comprendido en el paso 3 para cada uno de los elementos del conjunto de datos.

El esquema conceptual y los procesos de esta metodología se muestran a en las figuras 5.1 y 5.2, respectivamente.

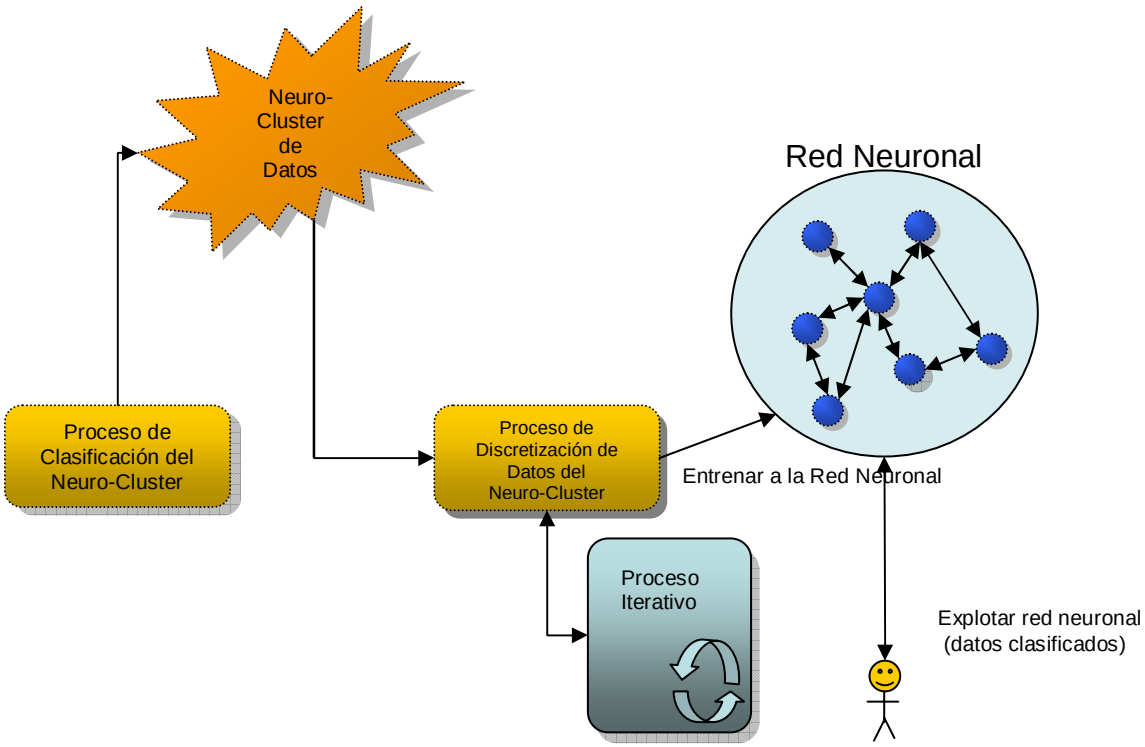


Figura 5.1. Sistema de clasificación de los datos de producción

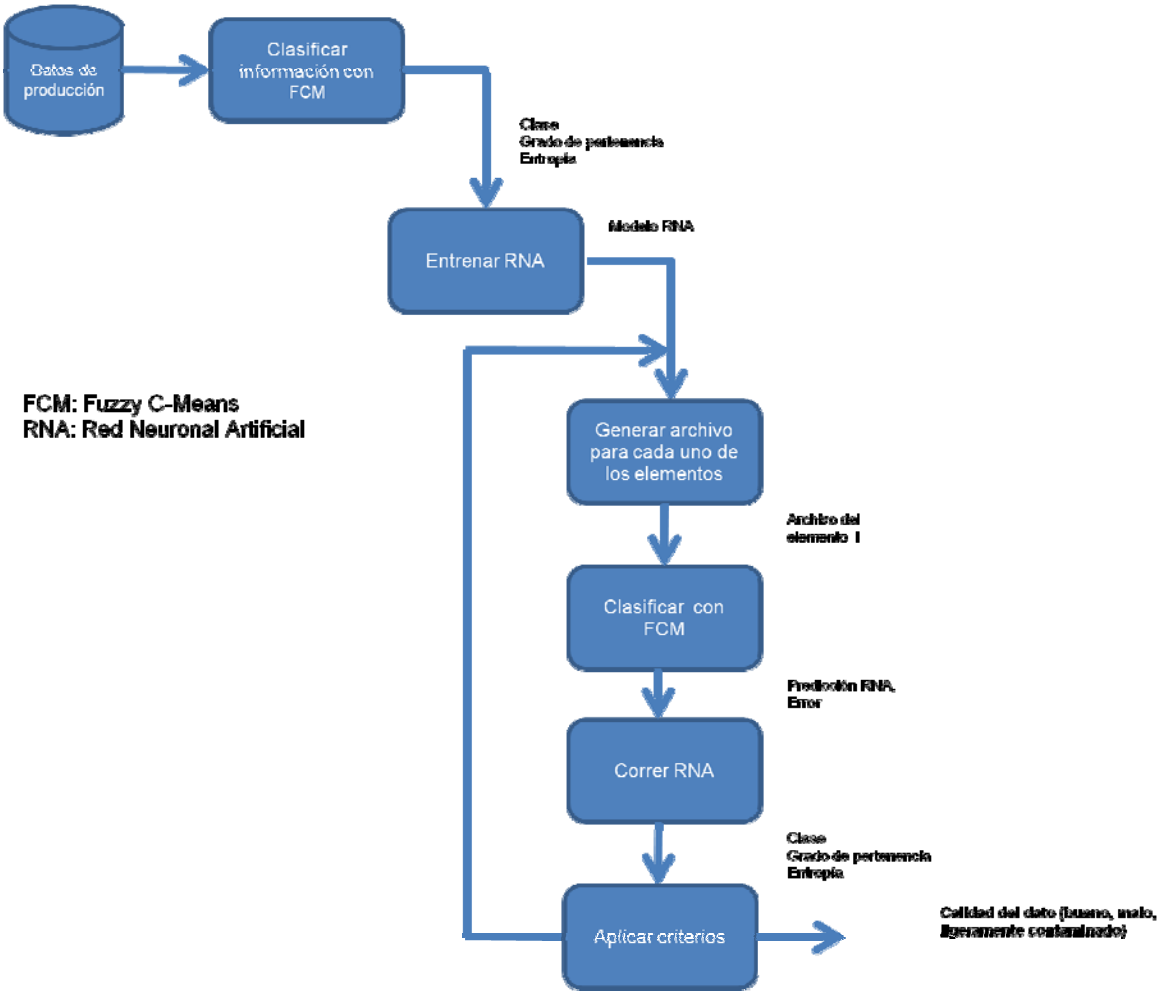


Figura 5.2. Tareas del método de validación de datos

Los criterios y restricciones para identificar la calidad de los datos se definen a continuación:

Un registro se considera “bueno”, cuando los valores de la curva de error y de la curva de entropía son cercanos a cero para un valor de la salida determinado, y además este valor pertenece a la clase dominante identificada por la curva de clase. En estos casos, la línea que representa la salida real, la línea que representa la predicción de salida de de la red neuronal artificial y el valor cero de la curva de error, deberán estar muy cercanos o traslaparse.

Un registro estará “ligeramente contaminado” cuando el valor de salida real o el de la salida de la red neuronal, sea cercano al valor cero en la curva de error, y una de las dos líneas de salida corresponda al valor mínimo de la curva de entropía. En estos

casos podría existir una diferencia significativa entre las dos líneas que representan las salidas.

Un registro se considera “malo” cuando no existe correlación entre la salida original, la salida de la red neuronal y la curva de error. Además, existe una diferencia notable entre las líneas, y no hay correlación con la curva de entropía.

El procedimiento iterativo antes mencionado requiere de un programa de cómputo que contemple los algoritmos de clasificación y los de redes neuronales artificiales, así como de módulos adicionales para conectar las salidas; además de otro que aplique los criterios de clasificación.

5.2. Requerimientos para la aplicación del método de validación del dato de producción

Previo a la aplicación del método propuesto, se deberá preparar la información. Durante esta etapa, el usuario deberá involucrarse completamente con el significado de los datos, ya que esto ayuda a construir un modelo más robusto. El proceso de preparación de datos consta de las fases siguientes:

a) Obtención de los datos.
b) Auditoría de los datos.
c) Enriquecimiento de los datos.
d) Búsqueda de sesgos en los datos.
e) Determinación de la estructura de los datos (modelo de datos).
f) Importación de datos.
g) Definición de parámetros de entrada, salida e identificadores de cada uno de los elementos del conjunto de datos.
h) Identificación y corrección de anomalías en los datos.
i) Construcción del modelo.

- Proceso de estimación de datos faltantes aplicando herramientas de la minería de datos.
- Manejo de datos fuera de tendencia.

j) Análisis de la información.

Cada una de las fases anteriores se describe a continuación.

a) Obtención de los datos.

Localizar los datos es el punto inicial y esencial para cualquier actividad de minería de datos. Aún cuando tengamos acceso a un almacén de datos donde se encuentra la información a utilizar, éstos pueden no estar en el formato necesario para su preparación. Generalmente los datos se crean con una estructura particular para reflejar algún punto de vista específico de la empresa, la cual podría alterar los resultados si no tenemos cuidado al momento de interpretarlos. Se puede tener una gran cantidad y tipos de fuentes de datos, y cada proyecto enfrenta diferentes retos, por lo que no será posible listarlos todos. Sin embargo, una solución es la importación de todas estas fuentes de datos a una base única de datos y organizarlos de tal manera que puedan presentarse en diferentes vistas, de acuerdo a los requerimientos particulares de cada proyecto.

b) Auditoría de los datos.

Una vez que los datos se han localizados, se tiene que determinar su estado y estimar si son apropiados para la construcción del modelo deseado. Se tiene que revisar sin son de cantidad suficiente y de buena calidad. Esto incluye el examinar el número de columnas y su contenido, los valores máximo y mínimo, el número de valores discretos y la aplicación de algunos otros criterios. Una vez que se haya asegurado la calidad y cantidad de los datos se podrá continuar. Como el objetivo del presente trabajo es validar datos, en este punto podría resultar confuso el término calidad de los datos. Por tanto, es importante aclarar que en este punto se está haciendo referencia a una validación general al conjunto de datos y no a cada uno de los elementos del conjunto de datos.

c) Enriquecimiento de los datos.

Si el resultado de auditar los datos nos indica que su estado actual no es adecuado para soportar una solución, es posible tratar de complementar el conjunto de datos. Agregar datos es una solución común para incrementar el contenido de la información. En la parte esencial de nuestro trabajo no lo haremos, ya que el objetivo es validar cada uno de los datos capturados durante el proceso de producción, para que posteriormente puedan utilizarse aquellos calificados como confiables (buenos y quizás los ligeramente contaminados). Una vez que se haya determinado la calidad de los datos, pueden enriquecerse tomándolos como referencia.

d) Búsqueda de sesgos en los datos.

Aún cuando algunos métodos automatizados puedan ayudar en la detección de sesgos en las muestras de datos, no sustituyen el razonamiento humano. Hay muchos métodos para el muestreo de datos, término que se refiere al proceso de extraer porciones pequeñas de datos contenidas en un gran conjunto de datos de tal forma que cada subconjunto refleje las relaciones contenidas en el conjunto completo. Esto puede ser difícil, ya que posiblemente las relaciones del conjunto no se vean reflejadas en un pequeño subconjunto, pero se debe trabajar en buscar que la muestra sea representativa.

e) Definición de la estructura de los datos.

Los metadatos son las estructuras de datos que describen las características de los datos de un sistema; se registran en ellos los nombres de los atributos, tipos de datos, y las restricciones de integridad de las relaciones especificadas en las bases de datos. El resultado de determinar la estructura es obtener los metadatos. Generalmente, los metadatos no permiten validar la calidad de los datos de una fuente de datos.

Normalmente los metadatos se representan con un modelo de datos o esquema conceptual, y generalmente su implementación física se realiza empleando como herramienta un motor de bases de datos, o un Sistema Administrador de Bases de Datos Relacionales (RDBMS por sus siglas en inglés). Existe toda una teoría

matemática, álgebra y cálculo relacional, en los que se basan este tipo de sistemas, sin embargo no son el motivo de este trabajo, por lo que no se discutirán en esta tesis; únicamente se mencionan por la importancia que tienen en nuestro modelo.

En esta etapa es muy importante la consistencia en los tipos de datos de las diferentes fuentes, y sus tipos correspondientes en las tablas físicas de la base de datos, para no tener problemas de pérdida de información.

f) Importación de datos

Una vez definido el esquema conceptual, e implementado el modelo físico en un sistema administrador de bases de datos, se procede a importar las diferentes fuentes de datos a las tablas físicas de la base de datos. El término importar datos se refiere a obtener datos a través de un programa informático, procedentes de otro programa. LA tarea inversa se conoce como exportar datos. Para realizar lo anterior se requiere cotejar el tipo de datos en los archivos fuentes y la factibilidad de importación, de tal manera que se disponga de los manejadores (drivers) que permitan la conexión del software administrador de bases de datos y las fuentes de datos (archivos en Excel, Access, etc.).

El éxito de esta actividad permitirá poder manipular y explotar la información conforme se requiera en la generación de nuestro modelo.

g) Definición de parámetros de entrada, salida e identificadores.

Con el fin de dar el sentido correcto en el análisis a los registros del conjunto de datos, es necesario indicar cuáles serán los atributos de entrada al sistema, cuáles los de salida y cuáles serán los identificadores para cada uno de los elementos del conjunto de datos.

Primero tenemos que indicar cuál o cuáles son los atributos que identificarán cada uno de los elementos. Para los datos del proceso de producción, estos atributos son el identificador del pozo (que puede ser el nombre del pozo, o número de pozo), además de la fecha de toma de la medición. El atributo de salida del sistema deberá ser la variable de análisis, que para el caso presente será el gasto de aceite. Los datos de entrada son aquellos atributos dentro del conjunto de datos, que influyen en el

comportamiento del atributo de salida (gasto de aceite). Los parámetros de entrada para el proceso de producción serán la presión en la tubería de producción, el diámetro del estrangulador, el gasto y presión de inyección del gas de bombeo neumático, el gasto de gas de la formación y la temperatura en la bajante.

h) Identificar y corregir anomalías en los datos.

En el contexto de limpieza de datos, los resultados del comportamiento y de la minería de datos, ayudan a identificar datos con patrones o relaciones inesperadas con otros atributos.

En la práctica, estos conjuntos de datos contienen varios tipos de elementos anómalos, que complican significativamente el problema de análisis; dos de las anomalías que suelen presentarse son:

- Datos fuera de tendencia,
- Datos faltantes.

La idea esencial en este proceso es hasta donde sea posible, el uso de una comparación sistemática y extensiva, que permita obtener la calidad necesaria para los resultados esperados.

Aún cuando sean pocas las anomalías, pueden tener una influencia desproporcionada en los resultados analíticos.

En algunos casos una comprensión preliminar de la naturaleza y fuente de este número reducido de anomalías, puede ser más valiosa que una compresión muy completa del resto del conjunto de datos.

Contradictoriamente, en casos donde estas anomalías no son de un interés inherente (por ejemplo errores por conversión), es importante no permitirles dominar los resultados del análisis.

- **Datos fuera de tendencia.**

Los datos fuera de tendencia son quizás el tipo más conocido y simple de anomalías de datos.

Un dato fuera de tendencia es un elemento anómalo con respecto al comportamiento visto en la mayoría de los otros elementos del conjunto de datos. Para convertir esta definición en un procedimiento para decidir cuando un elemento específico se encuentra fuera de tendencia, se requiere satisfacer los aspectos siguientes:

1. Una **caracterización** de la parte nominal (datos normales) del conjunto de datos.
2. Un criterio **cuantitativo** para decidir cuando el elemento en cuestión se encuentra en un conflicto significativo con la caracterización nominal.

La existencia de diferentes clases de comportamiento nominal nos lleva a la existencia de diferentes tipos de datos fuera de tendencia, y la existencia de diferentes criterios para la evaluación de desviaciones del comportamiento nominal, conduce a una variedad de algoritmos para detección de datos fuera de tendencia.

Una característica importante de este tipo de anomalía es que son valores extremos en la secuencia de datos. Si $x°$ es un dato fuera de tendencia y $x_a > x° \rightarrow x_a$ es un dato fuera de tendencia.

Probablemente el criterio más conocido para declarar que un valor de dato x_j en la secuencia $[x_k]$ es un dato fuera de tendencia es la regla: cualquier punto que se encuentre alejado más de tres desviaciones estándar de la media es un dato fuera de tendencia. Este procedimiento tiende a ejecutarse en forma deficiente en la práctica.

El dato de producción depende fuertemente de Δp, donde $\Delta p = \bar{p} - p_{wf}$, por lo que al momento de analizar el comportamiento del modelo, deberemos evaluar $q/\Delta p$, no solo q.

En la metodología propuesta se utilizarán curvas difusas conjuntamente con gráficas de regresión.

- **Datos Faltantes.**

Una anomalía extremadamente común en grandes conjuntos de datos es el dato faltante, correspondiente a valores de datos que deberían estar presentes en un

conjunto de datos, pero por varias razones, están ausentes. Una fuente común de datos faltantes son las fallas en los sistemas de medición.

Las consecuencias prácticas de los datos faltantes dependen de que proporción de datos están faltando y de su tipo. Los datos faltantes susceptibles a ignorarse, generalmente corresponden a la omisión de subconjuntos de datos seleccionados aleatoriamente. Puesto que la variabilidad de los resultados computados a partir de *n* datos normalmente decrece con el incremento de *n*, el efecto de datos faltantes susceptibles a ser ignorados es generalmente un incremento en la variabilidad relativa de los resultados que se deben obtener de un conjunto completo de datos. Por el contrario, generalmente los datos faltantes no susceptibles a ignorarse, corresponden a datos sistemáticamente faltantes de un conjunto de datos. La consecuencia de valores faltantes no susceptibles a ignorarse es más severa, ya que frecuentemente se introducen complicaciones en el resultado del análisis.

Para resolver esta anomalía puede aplicarse la metodología de *Parcheo Inteligente.* Esta metodología incorpora un sistema híbrido consistente de redes neuronales y algoritmos genéticos para identificar el mejor valor posible que se podría usar para parchar un hueco en el conjunto de datos.

Una manera de generar el modelo que permita predecir datos faltantes es partir de un conjunto de datos con información incompleta en algunos de los atributos. En éste se deberán definir los parámetros de configuración de una red neuronal artificial (razón de aprendizaje, momento y número de iteraciones) que en combinación con un algoritmo genético, el cual también deberá parametrizarse (número de generaciones, poblaciones, y porcentajes de cruzamiento, mutación e inversión), generarán el modelo de predicción.

i) Construcción del modelo.

Para generar el modelo de la red neuronal que se usará en este trabajo se requiere de tres subconjuntos de datos: datos de entrenamiento, datos de prueba y datos de calibración. Cada uno de estos subconjuntos deberá ser representativo y contener un detalle adecuado con el fin de soportar la solución esperada.

El proceso se puede resumir como sigue.

Dado un conjunto de datos**¡Error! Marcador no definido.**, se aplicarán las herramientas de modelado de minería de datos a través de una serie de pasos:

1. Reservar una porción de los datos disponible como un conjunto de evaluación para medir el comportamiento final.
2. Dividir el resto de los datos en datos de entrenamiento y datos de prueba.
3. Usando los datos de entrenamiento, generar modelos con varios grados de complejidad y suavidad usando diferentes arquitecturas de aproximación y parámetros de refinamiento. Esto se refiere a que para obtener el modelo que mejor represente el sistema buscado, se definirán una serie de escenarios con base en diferentes parámetros de configuración de la red neuronal artificial, como son el momento, el número de capas ocultas, el número de neuronas en las capas ocultas y la razón de aprendizaje.
4. Identificar los modelos que mejor funcionan con el conjunto de datos de prueba.
5. De los modelos que mejor funcionan, elegir el de menor complejidad y mayor suavidad.
6. Asegurar el comportamiento del modelo final usando los datos de evaluación. Entonces terminar.

La porción de los datos disponibles que se reserva para la evaluación final se llama frecuentemente conjunto de evaluación, la cual se usa para generar una prueba imparcial del modelo final antes de liberarse. Una cantidad suficiente de datos deberá reservarse para la prueba. Sin embargo, la mayoría de los datos se utilizarán para el entrenamiento y las pruebas. Aunque no hay proporciones fijas para dividir los datos, el conjunto de entrenamiento deberá ser más de la mitad (70% para entrenamiento y 30% para pruebas). Los datos en los dos conjuntos necesariamente serán aleatoriamente muestreados del conjunto de datos original.

Después, se generarán iterativamente formas parametrizadas ajustadas al conjunto de datos de entrenamiento, y se validarán utilizando los datos de pruebas. Los modelos generados deberán diferir en el grado de complejidad y de suavidad.

El comportamiento del modelo elegido puede evaluarse para calcular el error en el conjunto de evaluación. El desempeño del modelo deberá validarse por medio de un conjunto de datos de evaluación independiente de los datos de entrenamiento o de prueba, debido a que estos datos se usaron durante el desarrollo del modelo. En general, los errores en los datos de entrenamiento o de prueba serán más pequeños que aquellos que se obtengan en los de evaluación.

El cálculo del error de evaluación deberá realizarse solo después de que el modelado se ha terminado. Cualquier mejora adicional que se incorpore al modelo deberá emplear información obtenida utilizando el conjunto de datos de evaluación. Si se viola esta regla, el error del nuevo modelo generado con base en el conjunto de datos de evaluación será parcializado, como los errores con datos de entrenamiento y de pruebas.

j) Análisis de la información.

Esta etapa incluye el análisis estadístico básico y el avanzado. En el primero se llevará a cabo un análisis de las correlaciones existentes entre cada uno de los atributos del conjunto de datos, y se graficará las variaciones de los parámetros anteriores contra todos. Posteriormente se determinará en forma general cuáles son los atributos de mayor influencia en nuestra variable de salida, que es el gasto de aceite.

El propósito del análisis de datos es identificar errores e inconsistencias en grandes conjuntos de datos. Así, en esta fase se requiere un análisis cuidadoso y detallado. Diferentes herramientas pueden usarse en esta fase: análisis de metadatos, perfiles de datos, análisis estadístico, minería de datos, detección de datos fuera de la tendencia o atípicos, y tecnología de reconocimiento de patrones difuso (análisis combinatorio de los diferentes parámetros).

En el siguiente capítulo se describirá la aplicación de esta metodología.

CAPÍTULO 6. APLICACIÓN A UN CASO REAL.

6.1. Descripción del problema

La calidad de los datos de producción es un tema de interés general en la industria petrolera. En muchas ocasiones dichos datos son la única información disponible en cantidad suficiente en los campos maduros. Sin embargo, como mencionamos en un principio, la existencia de datos contaminados puede ocasionar fallas operacionales y además, conducir a la toma inadecuada de decisiones a nivel estratégico. Por esta razón es necesario tener certeza de la calidad de los datos de producción.

Con el fin de determinar la calidad de los datos de producción de un campo petrolero, y buscando probar la metodología de validación del dato de producción propuesta en el capítulo anterior, se realizaron las pruebas descritas a continuación con datos del Activo de Explotación Ku-Maloob-Zaap.

Inicialmente se intentó trabajar con la totalidad de los datos; sin embargo, se encontraron paulatinamente parámetros muy particulares para cada uno de los campos y para la formación explotada. Así, se encontró que los datos de Zaap en la formación a nivel Kimmeridgiano eran menos complejos, pero no había suficiente historia de explotación para el análisis. Finalmente se encontró que la dupla campo-formación con mayor información era el campo Ku en la formación del Cretácico Medio. La mayoría de estos pozos no son fluyentes, por lo que se tuvieron que incluir las variables correspondientes al proceso de Bombeo Neumático (BN).

Para la aplicación de la metodología se siguieron los pasos señalados en el capítulo anterior, documentándose en este capítulo el resultado de su aplicación.

Antecedentes del campo Ku

El campo Ku pertenece al Activo Integral Ku-Maloob-Zaap, localizado en la Región Marina Noreste, dentro de aguas territoriales del Golfo de México, a 105 Km al noroeste de Cd. del Carmen, tal como se muestra en Figura 6.1.

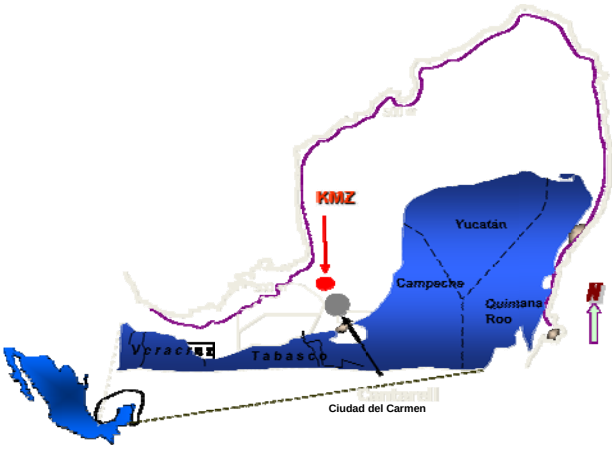


Figura 6.1. Localización del campo Ku-Maloob-Zaap

El campo Ku, uno de los más importantes en México en términos de la reserva remanente de hidrocarburos, fue descubierto en 1980 por el pozo Ha-1, productor en la brecha Paleoceno – Cretácico Superior. Más tarde, en marzo de 1981, fue terminado el pozo Ku-89, con una producción diaria de 22,000 barriles de aceite. Actualmente, la caída de presión ha dictado el uso de sistemas artificiales de producción, y el mantenimiento de presión por inyección de nitrógeno. Tal caída de presión ha creado un casquete de gas secundario. En este campo se produce aceite pesado, con una densidad que va desde los 12 a los 25° API.

6.2. Aplicación del método de validación del dato de producción a datos del campo Ku.

Este trabajo se desarrolló de la manera siguiente:

6.2.I. Preparación de la información:

a) Obtención de datos y estructura de la información (modelo de datos e importación).
b) Definición de parámetros de entrada, salida e identificadores.
c) Identificación y corrección de anomalías en los datos (estimación de datos faltantes y manejo de datos fuera de tendencia usando un modelo de red neuronal en conjunto con un algoritmo genético).
    i. Proceso de estimación de datos faltantes.
    ii. Identificación de datos fuera de tendencia.

d) Análisis de la información.

6.2.II. Método de validación de datos.

Las etapas anteriores se describen a continuación.

6. 2.I. Preparación de la información.

a) Obtención de datos

Se consiguió información de diferentes fuentes y disponible en varios formatos (Excel, Access, tablas de bases de datos), la cual se importó para su preparación y análisis a una base de datos en SQLServer (motor de bases de datos). Se diseñó un modelo de datos el cual se presenta en la Figura 6.2.

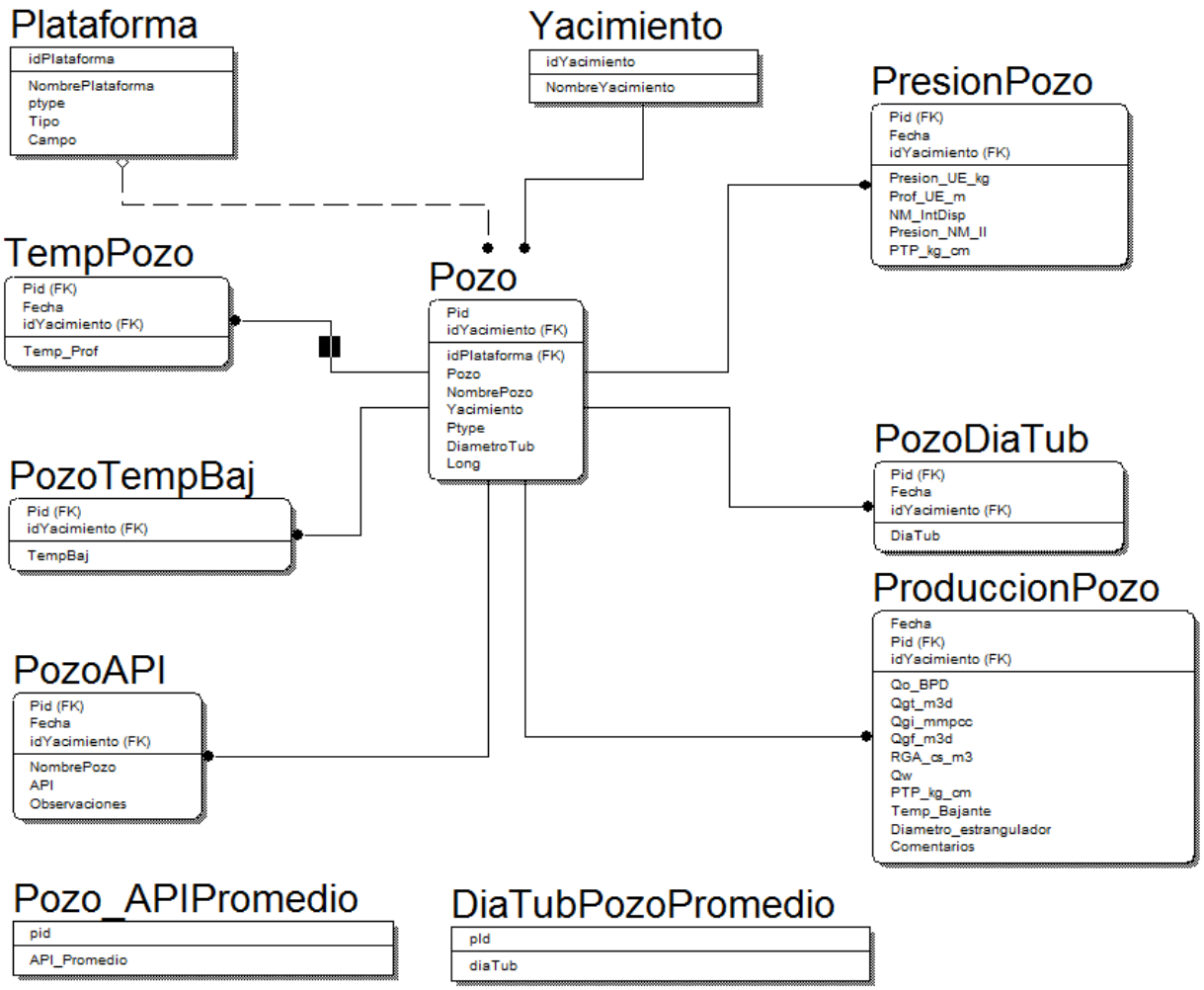


Figura 6.2. Modelo conceptual de la base de datos de producción de Ku-Maloob-Zaap .

En este modelo puede observarse la estructura de la información con la que se realizaron las pruebas. Cada una las cajas representa una entidad conceptual de información en nuestro sistema, y las líneas que las conectan representan las relaciones entre ellas. Básicamente estas entidades están definidas para representar el yacimiento, el pozo, la calidad la producción de hidrocarburos, la temperatura, y obviamente, la producción y las presiones en superficie. La relación de la entidad de información y su nombre en el modelo conceptual se presenta en la Tabla 6.1.

Tabla 6.1. Entidades conceptuales de la información del sistema.

Entidad de información	**Nombre en el modelo de datos**
Plataforma	Plataforma
Yacimiento	Yacimiento
Presión en el pozo	PresionPozo
Temperatura en el pozo	TempPozo
Pozo	Pozo
Diámetro de tubería del pozo	PozoDiaTub
Temperatura en la bajante	PozoTempBaj
Densidad del aceite producido en el pozo	PozoAPI
Densidad promedio del aceite producido en el pozo	Pozo_APIPromedio
Diámetro promedio de la tubería del pozo	DiaTubPozoPromedio

Cada una de las entidades tiene atributos o columnas que la caracterizan, lo cual puede observarse en la entidad Pozo, que tiene como atributos al identificador del pozo, su nombre, el yacimiento y plataforma a la cual pertenece, además del diámetro y longitud de tubería. Este tipo de modelos permite garantizar que la información que finalmente se guarde bajo este esquema, no sea redundante y tenga la integridad y consistencia básicas en el sistema, además de otros beneficios adicionales. Sin embargo, no puede garantizar la calidad de los datos que se registren en él.

Una vez definido e implementado el esquema conceptual, iniciamos la importación de datos. Así, se importaron las bases de datos en formato de Access (mdb) y en formato Excel (xls), a través de tareas en SQLServer. Una vez realizado lo anterior, se procedió

a efectuar un análisis preliminar a partir de consultas directas a la base de datos, con el fin de estudiar a nivel macro las irregularidades que pudieran contener. A continuación se mencionan las observaciones obtenidas.

- Existen 54 pozos en la base de datos cuyos nombres son los siguientes:
  KU-49, KU-5, KU-401, KU-84D, KU-67A, KU-1277, KU-87D, KU-82, KU-62A, KU-64, KU-46, KU-43, KU-27, KU-89, KU-61, KU-1001, KU-7, KU-21, KU-65, KU-83, KU-1272, KU-63D, KU-83D, KU-1275, KU-1297, KU-9, KU-1271, KU-44, KU-1293D, KU-69, KU-5D, KU-25, KU-47, KU-1299, KU-1278, KU-22, KU-288, KU-41, KU-66, KU-63, KU-84, KU-45, KU-45D, KU-26, KU-61D, KU-128D, KU-1292, KU-1291, KU-81, KU-1295, KU-1293, KU-42, KU-23, KU-87.
- Se observó en el análisis preliminar, que el pozo KU-21 tiene información incompleta de bombeo neumático (BN), ya que los gastos de gas totales, los gastos de gas de formación y los gastos de inyección de BN no concuerdan; además, las presiones de inyección tienen un valor nulo. Una muestra de los datos originales se presenta en el Apéndice A. Asimismo, se depuró la base de datos, eliminando las inconsistencias detectadas durante la importación de información.
- Se ejecutaron consultas adicionales a la base de datos con el fin de identificar los datos faltantes.

Con base en el análisis anterior se listan los comentarios siguientes:

1. Existen en la base de datos un total de 2,798 registros cuyo identificador es el nombre del pozo y la fecha de toma de la medición.
2. Existen 54 pozos distintos de las formaciones BP, KM, KI; en 5 plataformas distintas: KU-G, KU-A, KU-C, KU-F, KU-S y KU-I.
3. En algunos pozos con inyección de gas para BN se encontró que faltaba información de la presión de inyección. Esto se infiere porque al existir información de gasto de inyección de BN debería existir información de presión de inyección de BN.
4. Información faltante: presión en la bajante, presión en el separador, presión en la salida, gastos de inyección y presiones de inyección.

Como resultado de este análisis se realizaron las acciones siguientes:

1. Con el objetivo de aplicar el método de validación, se seleccionaron a través de una consulta a la base de datos, solo los datos de la formación Ku Cretácico Medio, cuyo resultado (información de 10 pozos diferentes) se exportó al archivo en formato de Excel llamado Ku-Cretácico-KM.xls. Se determinó seleccionar sólo una formación con el fin de que el modelo se aplicara a pozos con características semejantes. El resultado fue un conjunto de datos de 417 elementos.
2. Como los parámetros relacionados con el bombeo neumático son importantes en este estudio, también se integraron en el mismo. Para resolver el problema de datos de inyección de BN faltantes, se utilizó un método inteligente para estimarlos, el cual se describirá en el inciso (c).

b) Definición de parámetros de entrada, salida e identificadores.

Como se mencionó en el párrafo anterior, con el fin de dar sentido en nuestro análisis a los elementos del conjunto de datos, fue necesario indicar cuáles serían los atributos de entrada al sistema, cuáles los de salida y cuáles serían los identificadores para cada uno de los elementos del conjunto de datos. La base de datos (BD) de producción de Ku-Cretácico consistió de 20 parámetros, descritos en la Tabla 6.2.

Tabla 6.2. Parámetros incluidos en la base de datos y su rol dentro del desarrollo del modelo.

Descripción del parámetro	Nombre en la BD	Unidades	Rol
Nombre del Pozo	Pozo		Identificador
Campo al que pertenece	Campo		No usado
Formación	Formación		No usado
Plataforma	Plataforma		No usado
Fecha de toma de medición	Fecha	dd/mm/aaaa	Identificador
Diámetro de estrangulador	Estrangulador_inch	pulgadas	Entrada
Presión en la tubería de producción	PTP_kg_cm2	kg/cm^2	Entrada
Presión de inyección de bombeo neumática	Piny_kg_cm2		Entrada
Presión en bajante	Pbajante_kg_cm2	kg/cm^2	Entrada
Presión en el separador	Psep_kg_cm2	kg/cm^2	Entrada
Presión a la salida	Psalida_kg_cm2	kg/cm^2	Entrada
Temperatura en el separador	TSep_DegC	°C	Entrada
Gasto de inyección de BN	InyeccionBN_MMpc_d	MMpc/d	Entrada
Gasto de aceite	Qo_bbl_d	Bbl/d	Salida
Gasto de gas total (inyección BN + gas de formación)	Qg_MMpc_d	MMpc/d	Entrada
Temperatura en la bajante	Tbajante_DegC	°C	No usado
Gasto de gas de la formación	QgForm_MMpc_d	MMpc/d	Entrada
RGA	RGA_m3_m3	m^3/m^3	No usado
RGIL	RGIL_m3_m3	m^3/m^3	No usado
Observaciones	Observaciones		No usado

c) Identificación y corrección de anomalías en los datos

Las anomalías de los datos detectadas y manejadas en la información disponible fueron los datos faltantes y aquellos fuera de tendencia. Como se mencionó en párrafos anteriores, el proceso de estimación de datos faltantes se resolvió utilizando herramientas de minería de datos.

i) Proceso de Estimación de datos faltantes

Para estimar los datos que faltaban en la columna de presión de inyección de BN, se utilizó un proceso de parcheo inteligente. Este proceso incorpora un sistema híbrido

consistente de una red neuronal artificial y un algoritmo genético, para identificar el mejor valor posible que puede usarse, con el fin de llenar un hueco en el conjunto de datos. Para la aplicación del modelo, se utilizaron las herramientas que forman parte del software *IDEA (Intelligent Data Evaluation & Analysis)*1.

Para realizar esta estimación, primero se efectuó una prueba con un archivo con la misma estructura, pero con datos alterados intencionalmente. El objetivo era determinar la validez del modelo de estimación que se generó, para después aplicarlo al total de los datos con registros con presión de inyección de BN faltante.

Para seleccionar el conjunto de datos a utilizar en la primera etapa de esta prueba, se partió de la información siguiente: de un total de 417 registros, se tuvieron 286 registros incompletos, es decir un 68.6% de registros incompletos. El parámetro de presión de inyección de BN (no alterado intencionalmente) tuvo 211 registros incompletos.

Se dispuso de 134 registros con datos de superficie casi completos. Una muestra de estos registros se incluye en el Apéndice A. Solo faltaban algunos datos de temperatura en la bajante, razón por la que no se consideró este atributo como parámetro de entrada.

Antes de iniciar el proceso, se eliminaron intencionalmente las presiones de inyección de gas de BN de tres elementos del conjunto de datos de tres pozos diferentes. Estos valores se presentan en la Tabla 6.3.

Tabla 6.3. Valores de presión de inyección de BN alterados intencionalmente.

Presión de inyección de BN eliminado	**Valor [kg/cm^2]**
PInyBN1	56.70
PInyBN2	66.00
PinyBN3	56.70

En la Tabla 6.4 se presentan los registros completos correspondientes a la Tabla 6.3.

Tabla 6.4. Datos adicionales de los registros alterados intencionalmente.

Pozo	Fecha	Estrangulador_inch	PTP_kg_cm2	Piny_kg_cm2	InyeccionBN_MMpc_d	Qo_bbl_d	Qg_MMpc_d	QgForm_MMpc_d
KU-1001	9/19/2003	3.88	17.40	56.70	2.51	6328.00	4.97	2.46
KU-89	3/20/2003	3.25	14.00	66.00	1.71	5372.00	3.80	2.09
KU-67A	4/4/2002	3.25	12.00	56.70	1.25	6486.00	4.21	2.97

Posteriormente, se importaron los 134 registros al módulo de parcheo inteligente del software IDEA[1], y se seleccionaron los parámetros de entrada, de salida e identificadores, de acuerdo a la

Entre los 134 registros anteriores, se encontró que el pozo Ku-63 tenía un dato en el que a una presión de inyección de BN de 66 kg/cm^2, el gasto de inyección era cero, el cual se modificó a un valor nulo con el fin de que se estimara a través del modelo generado. En ese momento se tenían tres valores de presión de inyección alterados intencionalmente y un valor llevado a nulo por no estar disponible y considerarse sin sentido físico.

Se utilizó el coeficiente de correlación R^2 como un indicador de la bondad de ajuste del modelo a los datos. En la ecuación 6.1 la variable Q_o representa el gasto de aceite.

$$\boldsymbol{R^2 = 1 - \frac{\sum_1^n \left[Qo_{real} - Qo_{estimado}\right]^2}{\sum_1^n [Qo_{real} - \overline{Qo}]^2}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (6.1)$$

Con el fin de determinar la configuración de los parámetros de la red neuronal artificial (RNA) y del algoritmo genético (AG), que permitieran generar el modelo óptimo para la estimación de datos de presión de inyección de BN, se prepararon diferentes escenarios. De esta forma, se encontró que el valor no disponible (nulos) de gasto de inyección de BN no se obtenían buenos resultados, ya que los valores de R^2 eran muy bajos. Por tanto, se dejó el valor cero que originalmente tenía y los resultados mejoraron. Los resultados finales se presentan en la Tabla 6.5. En la primera columna aparece el índice o identificador del escenario. En las siguientes aparecen los parámetros de configuración de la RNA, los parámetros y probabilidades del algoritmo genético y R^{2,} tanto para datos completos como para datos completos más parchados.

Las últimas columnas muestran un comparativo entre los tres valores de presión de inyección de BN alterados intencionalmente (valor medido y valor estimado).

Tabla 6.5. Resultados de las pruebas de estimación de datos faltantes, los valores de presión están en kg/cm$^{2.}$

No. De prueba	Parámetros de la Red Neuronal Artificial			Parámetros del Algoritmo Genético:		Probabilidades del Algoritmo Genético:			R^2 (Real-Predicción RNA)		Solo registro parchados	PInyBN1		PInyBN2		PInyBN1	
	Razón de aprendizaje	Momento	Iteraciones	Generaciones	Poblaciones	Cruzamiento	Mutación	Inversión	Datos Completos	Todos		Medido	Estimado	Medido	Estimado	Medido	Estimado
1	0.30	0.80	200	10	20	60	2	5	0.99	0.97	0.95	56.70	68.01	66.00	43.65	56.70	60.66
2	0.30	0.80	220	10	20	60	2	5	0.99	0.99	0.96	56.70	72.00	66.00	52.15	56.70	59.53
3	0.30	0.80	240	10	20	60	2	5	0.99	0.99	0.84	56.70	68.60	66.00	56.69	56.70	72.00
4	0.30	0.80	260	10	20	60	2	5	0.99	0.98	0.97	56.70	53.86	66.00	52.16	56.70	70.30
5	0.30	0.80	280	10	20	60	2	5	0.99	0.99	0.60	56.70	72.00	66.00	64.63	56.70	69.73
6	0.30	0.80	300	10	20	60	2	5	0.99	0.99	0.72	56.70	65.76	66.00	65.20	56.70	61.80
7	0.30	0.80	400	10	20	60	2	5	0.99	0.99	0.85	56.70	64.06	66.00	7.94	56.70	72.00
8	0.30	0.80	1000	10	20	60	2	5	1.00	1.00	0.99	56.70	58.39	66.00	65.76	56.70	42.52
9	0.30	0.80	280	10	20	60	2	5	0.99	0.99	0.91	56.70	0.57	66.00	36.28	56.70	68.60
10	0.40	0.80	280	10	20	60	2	5	0.99	0.99	0.89	56.70	70.87	66.00	49.32	56.70	65.76
11	0.50	0.80	280	10	20	60	2	6	1.00	0.99	0.89	56.70	72.00	66.00	71.43	56.70	68.03
12	0.70	0.80	280	10	20	60	2	6	1.00	0.99	0.88	56.70	18.14	66.00	72.00	56.70	72.00
13	0.80	0.80	280	10	20	60	2	6	0.99	0.99	0.53	56.70	15.31	66.00	65.76	56.70	65.76
14	0.50	0.60	280	10	20	60	2	6	0.98	0.98	0.79	56.70	38.88	66.00	28.35	56.70	71.43
15	0.50	0.70	280	10	20	60	2	6	0.99	0.98	0.39	56.70	30.05	66.00	71.43	56.70	39.69
16	0.50	0.80	280	10	20	60	2	5	0.99	1.00	0.90	56.70	35.58	66.00	54.99	56.70	68.03
17	0.50	0.90	280	10	20	60	2	5	1.00	0.99	0.99	56.70	35.72	66.00	39.69	56.70	52.72

Se graficaron los diferentes valores de R^2 para dos conjuntos: 1) datos completos, 2) datos completos más parchados; y para cada uno de los escenarios de los parámetros del Algoritmo Genético (AG) y de la Red Neuronal artificial (RNA). En la Figura 6.3 se presenta el comportamiento del coeficiente de correlación R^2 para diferentes números de iteraciones. Aparecen dos curvas, la primera corresponde al conjunto de datos completos y la segunda al de datos completos más parchados. El punto en el que ambas curvas se juntan y para un valor de R^2 cercano a 1, corresponde al mejor resultado, obtenido para un número máximo de 1000 iteraciones. De aquí obtuvimos el número óptimo de las mismas a configurar en la RNA.

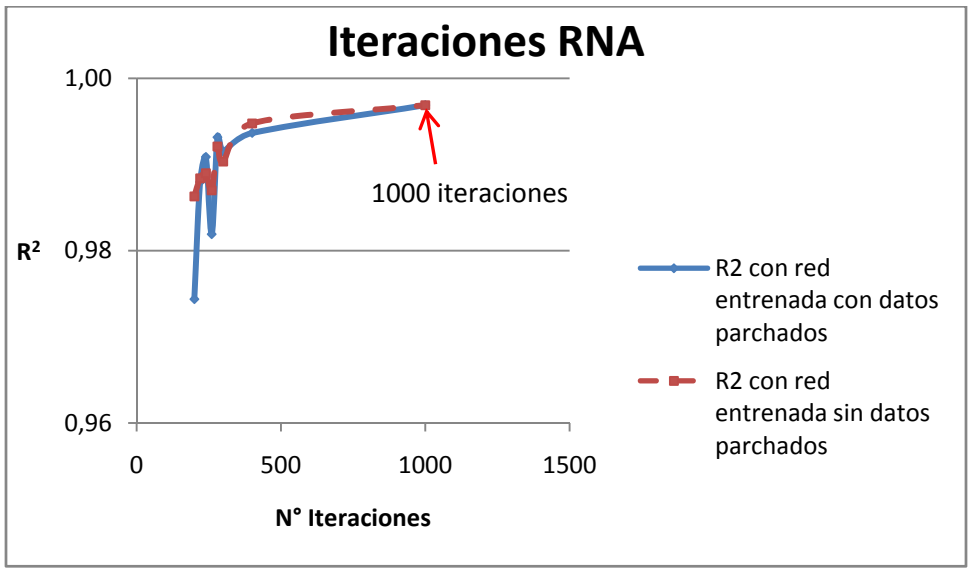


Figura 6.3. Comportamiento de R^2 para diferentes valores de número de iteraciones.

Para determinar el valor óptimo de la razón de aprendizaje de la RNA, se aplicó el mismo criterio considerado para el número de iteraciones. En la Figura 6.4 se presentan dos curvas, la de datos completos, y la de datos completos más los parchados. El punto de mayor valor de R^2 en el que coinciden las dos curvas, corresponde a una razón de aprendizaje de 0.3.

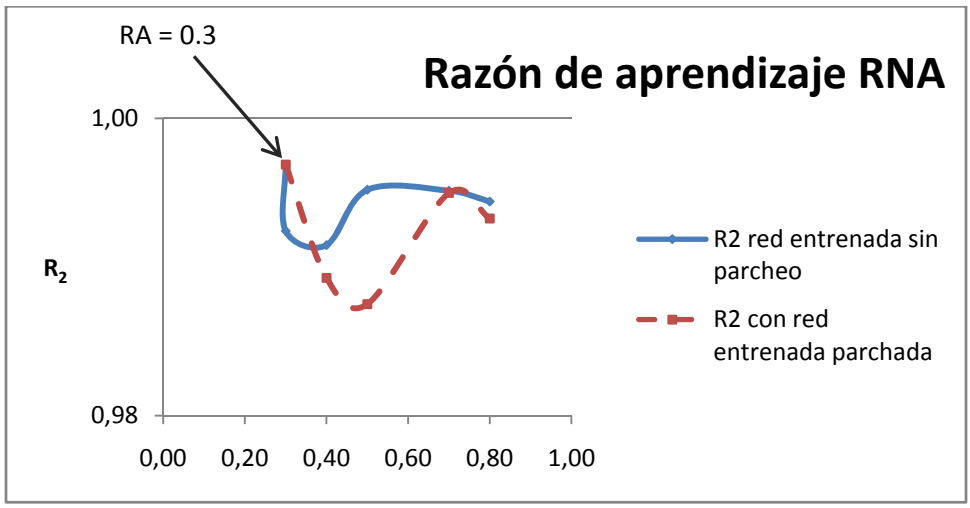


Figura 6.4. Comportamiento de R^2 para 280 iteraciones y para diferentes valores de razón de aprendizaje (RA).

Una vez que se obtuvieron el número óptimo de iteraciones y la razón de aprendizaje de la RNA, se procede a estimar el número óptimo para el momento de la RNA. El procedimiento es el mismo seguido para los dos parámetros anteriores; en la Figura 6.5 se observan los resultados calculados. Para obtener estos valores se fijaron los valores de la razón de aprendizaje y el número de iteraciones en 0.3 y 1000, respectivamente.

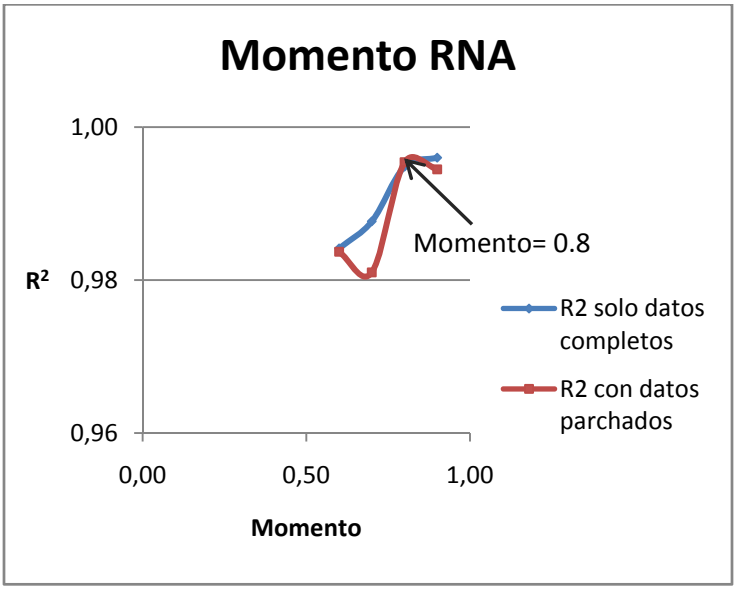


Figura 6.5. Comportamiento de R^2 con razón de aprendizaje de 0.3 y 1000 iteraciones.

Después de determinar los parámetros óptimos de configuración de la RNA, se utilizó el procedimiento previo para calcular los parámetros del algoritmo genético. El resultado

se muestra en la Tabla 6.5. Finalmente se obtuvo el comparativo final de los valores de R^2 para el conjunto de datos completos y el conjunto de datos completos más los parchados, el cual se presenta en la Figura 6.6. El punto en el que ambas curvas se juntan y el valor de R^2 es más cercano a 1, corresponde al escenario 8, por lo que es el que presenta los mejores resultados. En la Tabla 6.5 puede observarse que el escenario 8 tiene un coeficiente de correlación casi perfecto (0.9973 y 0.99582, redondeados a 2 decimales) tanto para los datos completos como para los datos completos más parchados. Asimismo, en la columna de la Tabla 6.5 etiquetada como "Solo registros parchados", se puede observa que el valor de R^2 es de 0.99. Este valor es el coeficiente de correlación entre los valores de salida, Q_o, de los registros que fueron parchados, antes y después de su estimación. Los valores del escenario 8 son los que deben configurarse en la RNA y en el AG. Estos valores se emplearon en el software IDEA[1] para utilizar el módulo de parcheo inteligente.

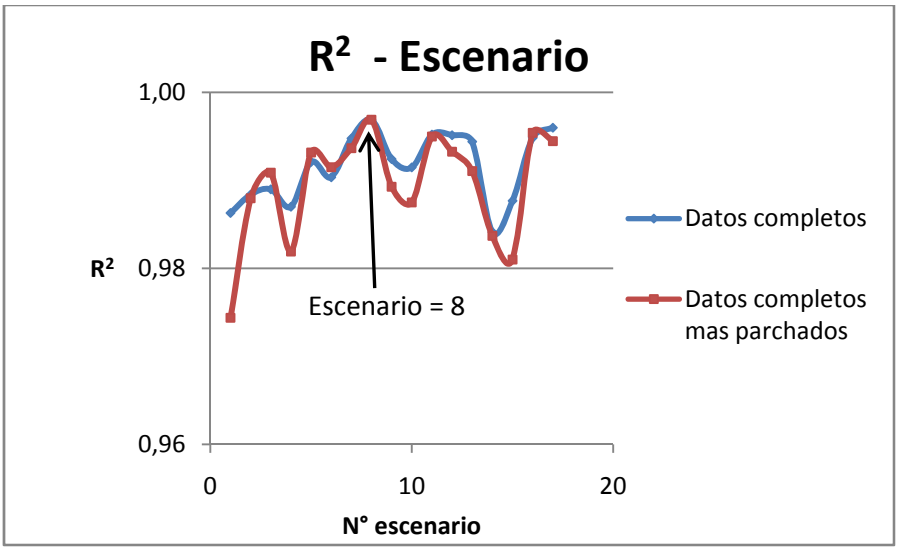


Figura 6.6. Selección del escenario con mejor resultado.

En la Figura 6.7 se muestra la configuración de parámetros de la RNA y el AG, con el software IDEA[1]. En la parte inferior aparecen los valores configurados. Además, en la tabla principal incluida en la misma figura, podemos observar el valor estimado del registro con índice 5, que fue uno de los alterados intencionalmente.

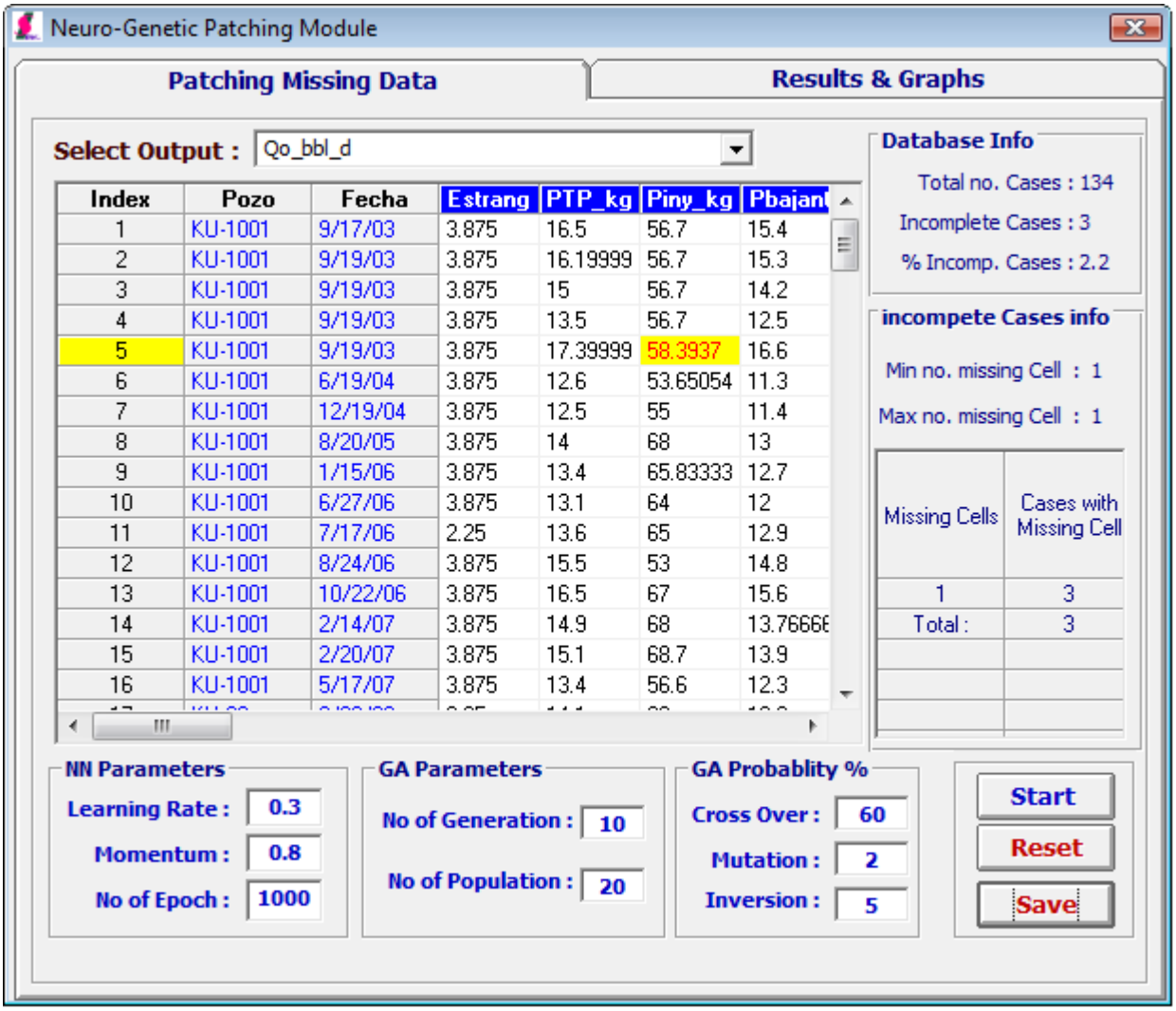


Figura 6.7. Configuración de parámetros de RNA, y AG en software IDEA[1].

En la Figura 6.8 se puede observar un comparativo entre los valores reales y los valores estimados por el modelo. La curva con rombos muestra los valores reales y la curva con triángulos presenta la predicción. Los valores de R^2 incluidos en la parte inferior izquierda son los que se tienen en la Tabla 6.5, solo que con mas cifras decimales.

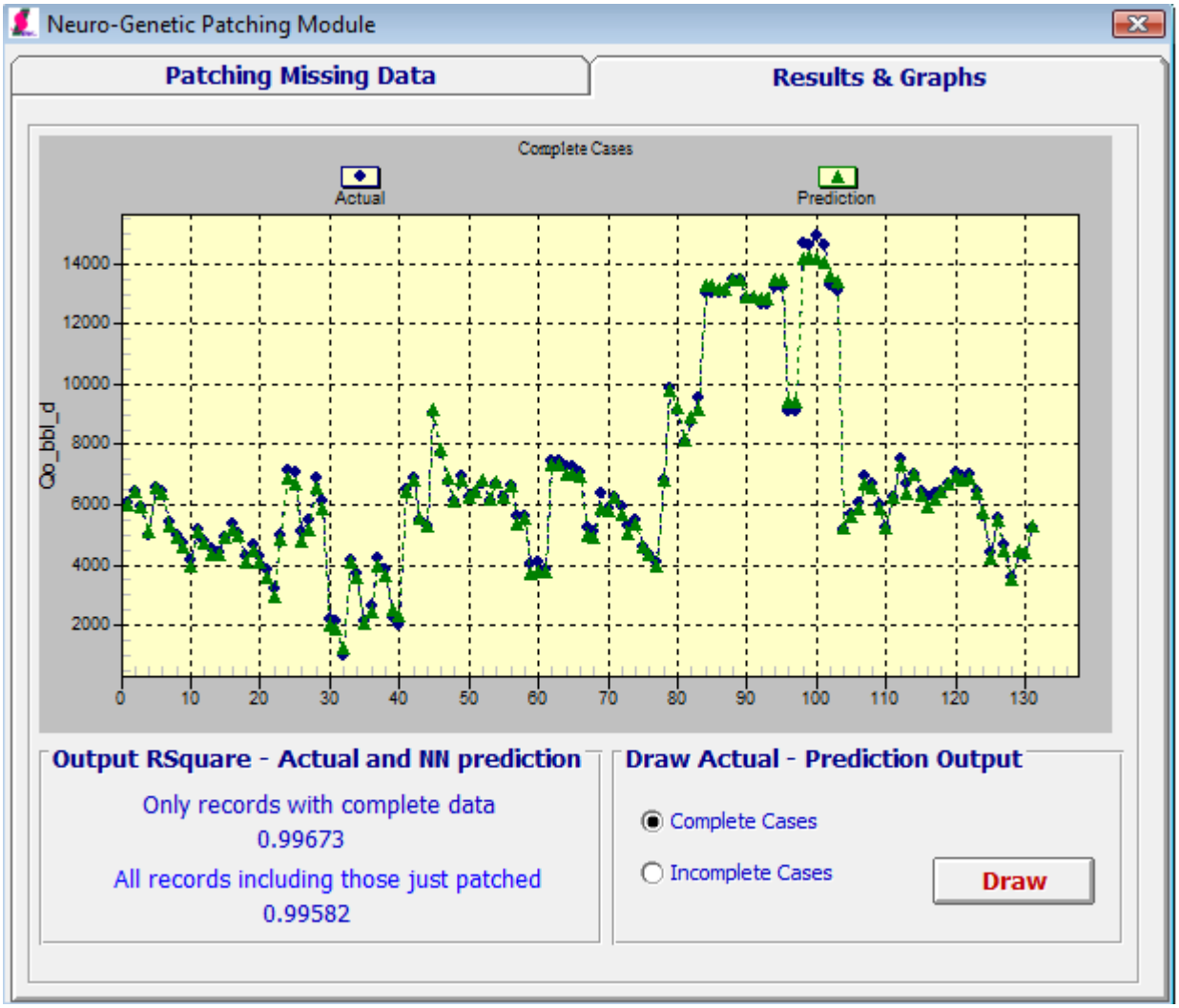


Figura 6.8. Resultado de la estimación de datos faltantes.

En la Figura 6.9 se presentan los resultados para los tres registros con información faltante; se puede observar como la correlación entre los valores de Q_o reales y estimados es de 0.99312.

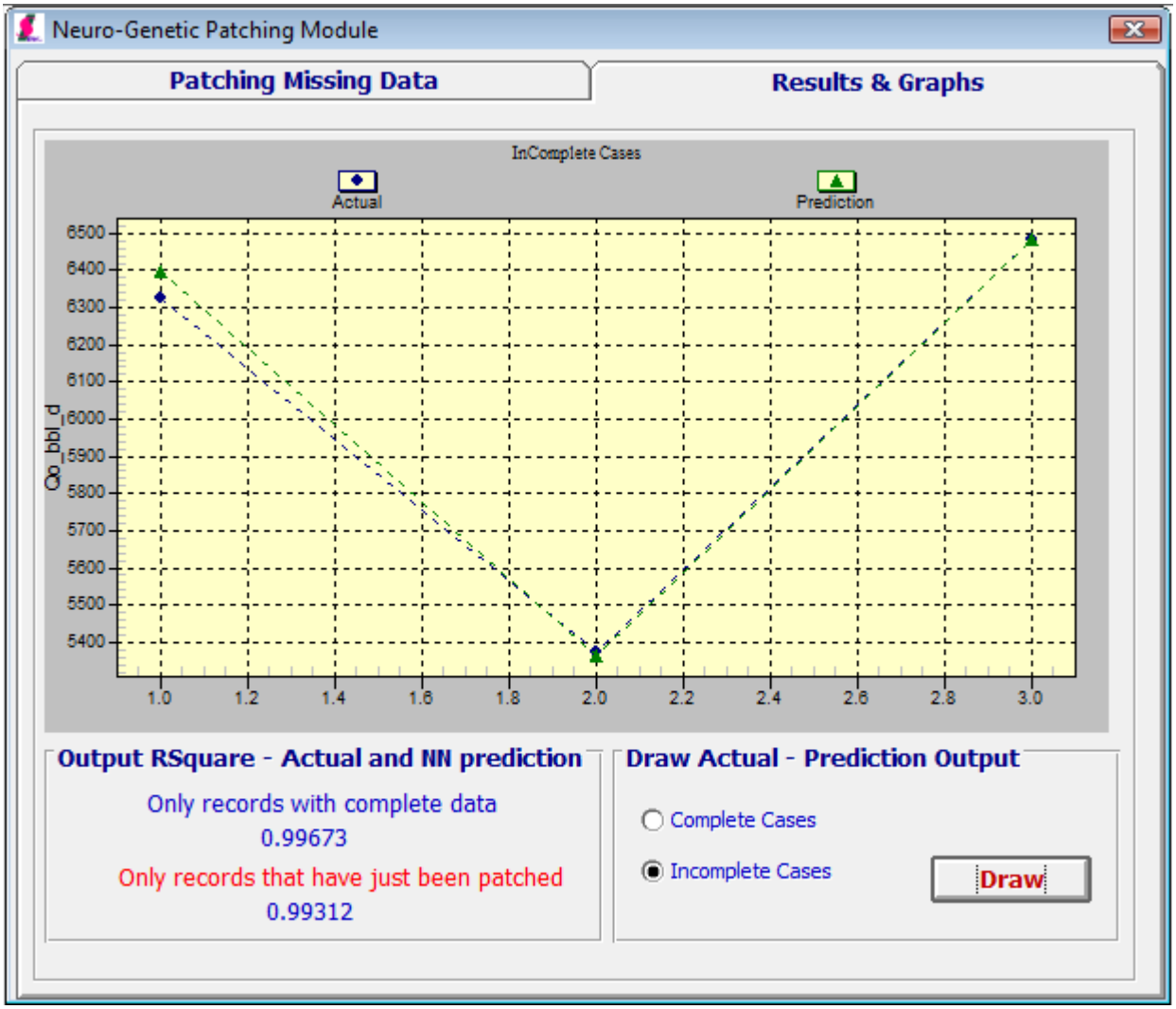


Figura 6.9. Resultado obtenido para el conjunto de datos incompletos.

En la Tabla 6.6 se muestra un comparativo de los valores reales contra los valores estimados de la presión de inyección de BN.

Tabla 6.6 Comparativo de los valores reales y estimados de la presión de inyección de BN.

Pozo	Presión de inyección de BN Eliminado	Valor real [kg/cm^2]	Valor estimado [kg/cm^2]
KU-1001	PInyBN1	56.70	58.39
KU-89	PInyBN2	66.00	65.76
KU-67A	PinyBN3	56.70	42.52

Con base en los resultados anteriores, se concluye adecuadamente el proceso de estimación de datos faltantes para el conjunto en el que se eliminó información intencionalmente, y el modelo generado se podrá utilizar para la estimación de los datos de presión de inyección faltante en el resto de los registros.

Aun cuando el proceso de estimación de datos faltantes fue exitoso, en esta aplicación no se utilizará, ya que se desea partir de los datos reales para validar su confiabilidad. Por esta razón se retoma al archivo original de los 134 registros completos, con el fin de continuar con el análisis estadístico básico. Se determinó nuevamente el identificador de cada registro, así como parámetros de entrada y parámetros de salida. Inicialmente se conservaron todos los parámetros con el fin de analizar cada uno y evaluar cuales son los de mayor influencia en el proceso.

ii) Identificación de datos fuera de tendencia

Se identificaron los registros con datos fuera de tendencia por medio de un análisis de regresión. Para realizar esta tarea se utilizó nuevamente el software IDEA[1]. Se graficaron todos los parámetros en el eje *X* contra Q_o en el eje *Y*. Visualmente se obtuvieron cuatro registros, los cuales se marcaron para interpretaciones futuras. En la Figura 6.10 se muestra el comportamiento de la presión en la tubería de producción p_{tp}, contra el gasto de aceite, Q_o; el punto identificado con un cuadro corresponde al dato fuera de tendencia correspondiente al pozo Ku-47.

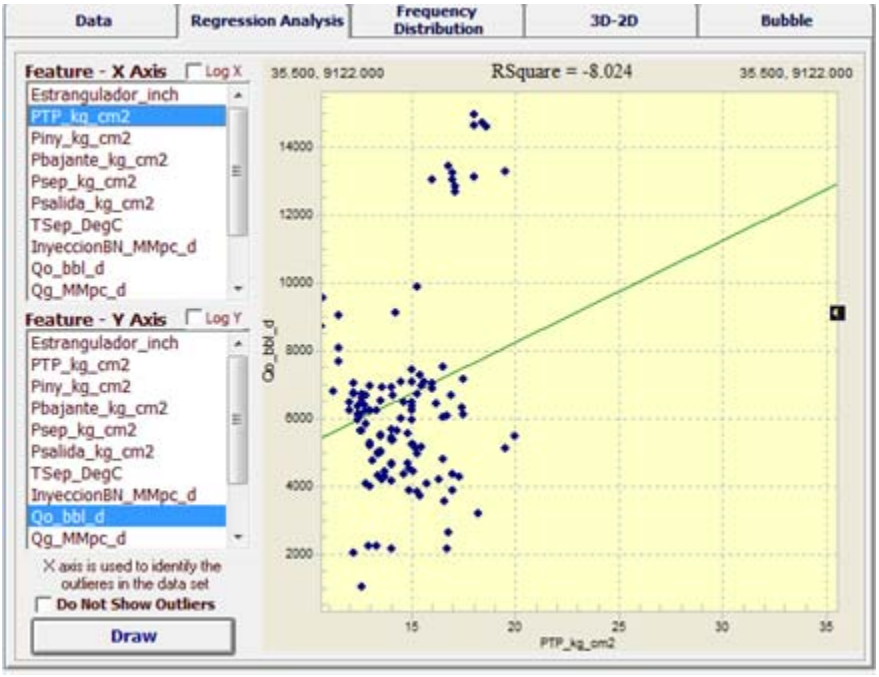


Figura 6.10. Análisis de regresión, p_{tp} contra Q_o

En la Figura 6.11 se muestra el comportamiento de la presión en la bajante, $p_{bajante}$ contra el gasto de aceite, Q_o. El punto indicado con un cuadro corresponde a los valores fuera de tendencia, y también corresponden al pozo Ku-47. El objetivo no fue encontrar alguna correlación, sino observar los datos fuera de tendencia.

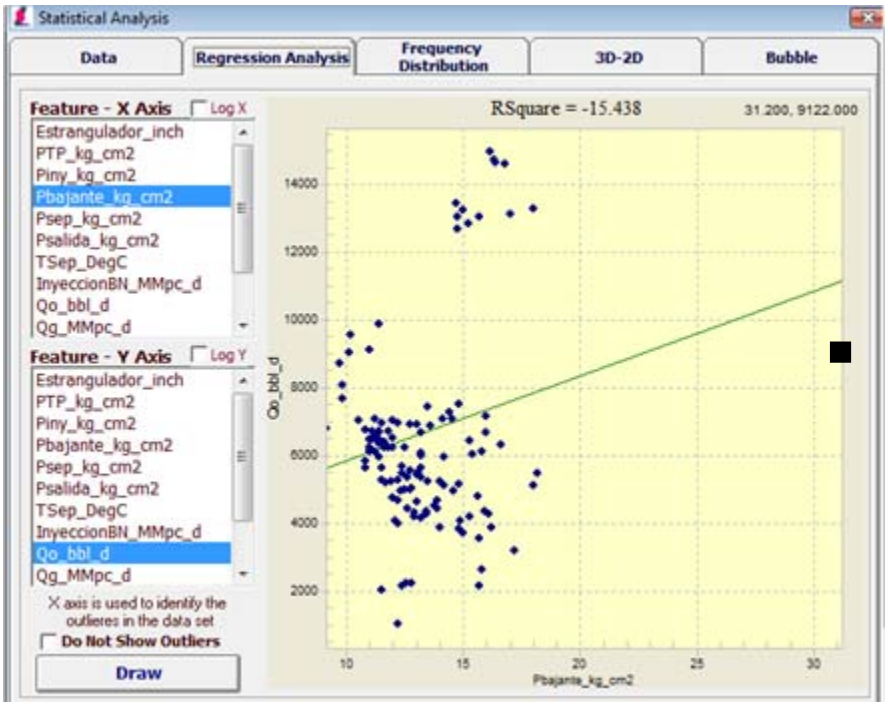


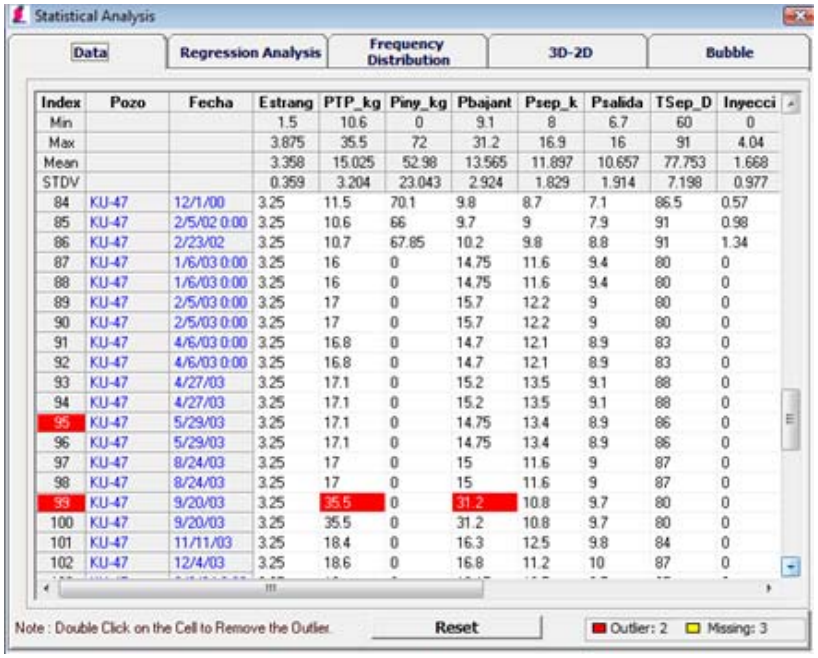
Figura 6.11. Análisis de regresión, $p_{bajante}$ contra Q_o

Estos cuatro registros pertenecen al pozo Ku-47. Los atributos de presión en la tubería de producción salieron de la tendencia en las mediciones tomadas el 29 de marzo del 2003, y en el gasto de gas de las mediciones tomadas el 20 de septiembre del mismo año. Estos registros se eliminaron, quedando el conjunto de datos con 130 registros. En la Tabla 6.7 se presentan los resultados del análisis de los datos fuera de tendencia, en donde se sombrearon aquellos que fueron eliminados, debido a que las presiones en la tubería de producción y los gastos de gas eran atípicos con respecto al conjunto de datos completo. Se puede observar que los registros aparecen duplicados, porque así estaban en la base de datos original.

Tabla 6.7. Datos fuera de tendencia se presentan sombreados.

Pozo	Fecha de medición	Diámetro estrangulador	PTP	PinyBN	Pbaj	Psep	PSali	TSep	QinyBN	Qo	QgTot	QgFor
KU-47	29/05/2003	3.25	17.1	0	14.75	13.4	8.9	86	0	12662	22.05	22.05
KU-47	29/05/2003	3.25	17.1	0	14.75	13.4	8.9	86	0	12662	22.05	22.05
KU-47	20/09/2003	3.25	35.5	0	31.2	10.8	9.7	80	0	9122	14.15	14.15
KU-47	20/09/2003	3.25	35.5	0	31.2	10.8	9.7	80	0	9122	14.15	14.15

En la Figura 6.12 se muestran los resultados obtenidos con el software IDEA[1]; se pueden observar las diferencias entre las presiones identificadas como fuera de tendencia con respecto a las del resto de los datos del pozo Ku-47.



Index	Pozo	Fecha	Estrang	PTP_kg	Piny_kg	Pbajant	Psep_k	Psalida	TSep_D	Inyecci
Min			1.5	10.6	0	9.1	8	6.7	60	0
Max			3.875	35.5	72	31.2	16.9	16	91	4.04
Mean			3.358	15.025	52.98	13.565	11.897	10.657	77.753	1.668
STDV			0.359	3.204	23.043	2.924	1.829	1.914	7.198	0.977
84	KU-47	12/1/00	3.25	11.5	70.1	9.8	8.7	7.1	86.5	0.57
85	KU-47	2/5/02 0:00	3.25	10.6	66	9.7	9	7.9	91	0.98
86	KU-47	2/23/02	3.25	10.7	67.85	10.2	9.8	8.8	91	1.34
87	KU-47	1/6/03 0:00	3.25	16	0	14.75	11.6	9.4	80	0
88	KU-47	1/6/03 0:00	3.25	16	0	14.75	11.6	9.4	80	0
89	KU-47	2/5/03 0:00	3.25	17	0	15.7	12.2	9	80	0
90	KU-47	2/5/03 0:00	3.25	17	0	15.7	12.2	9	80	0
91	KU-47	4/6/03 0:00	3.25	16.8	0	14.7	12.1	8.9	83	0
92	KU-47	4/6/03 0:00	3.25	16.8	0	14.7	12.1	8.9	83	0
93	KU-47	4/27/03	3.25	17.1	0	15.2	13.5	9.1	88	0
94	KU-47	4/27/03	3.25	17.1	0	15.2	13.5	9.1	88	0
95	KU-47	5/29/03	3.25	17.1	0	14.75	13.4	8.9	86	0
96	KU-47	5/29/03	3.25	17.1	0	14.75	13.4	8.9	86	0
97	KU-47	8/24/03	3.25	17	0	15	11.6	9	87	0
98	KU-47	8/24/03	3.25	17	0	15	11.6	9	87	0
99	KU-47	9/20/03	3.25	35.5	0	31.2	10.8	9.7	80	0
100	KU-47	9/20/03	3.25	35.5	0	31.2	10.8	9.7	80	0
101	KU-47	11/11/03	3.25	18.4	0	16.3	12.5	9.8	84	0
102	KU-47	12/4/03	3.25	18.6	0	16.8	11.2	10	87	0

Figura 6.12 Datos fuera de tendencia detectados por el software IDEA[1]

d) Análisis de la información

Con el fin de establecer las relaciones entre los diferentes parámetros, se realizó un análisis estadístico básico, el cual permitió identificar correlaciones entre atributos y entre grupos de atributos. Es conveniente recordar que el conjunto de datos analizado corresponde a diez diferentes pozos de la formación Ku-KM. En la Tabla 6.9 se presentan las relaciones encontradas. En las dos primeras columnas se presentan los parámetros que se están comparando y en la tercera el coeficiente de correlación entre ellos. Los parámetros en los que existe una correlación se identifican con una sombra. Puede observarse, como se esperaría, que están correlacionados con el gasto de aceite los siguientes parámetros: presión de inyección de BN, gasto de gas (total y de formación), y la relación gas aceite. Las presiones de superficie (bajante y de tubería de producción) están altamente correlacionadas. Asimismo, el gasto de gas (el de formación y el total) y la relación gas aceite se correlacionan con la presión de inyección de BN. Finalmente, los gastos de gas y la relación gas aceite, están altamente correlacionados.

Tabla 6.8. Resultado de la correlación entre atributos.

Parámetro X	*Parámetro Y*	R^2
Diámetro Estrangulador	*Qo*	*-261.154*
Presión TP	*Qo*	*-8.024*
Presión Iny BN	*Qo*	*0.555*
Presión Bajante	*Qo*	*-15.438*
Presión Separador	*Qo*	*-70.396*
Presión Salida	*Qo*	*-2.999*
Temp Separador	*Qo*	*-0.762*
QInyBN	*Qo*	*-0.073*
Qg Total (BN+Formación)	*Qo*	*0.699*
QgForm	*Qo*	*0.761*
RGA	*Qo*	*0.423*
RGIL	*Qo*	*0.435*
Qo	*DiámetroEstrang*	*Menor a cero*
Presión TP	*DiámetroEstrang*	*Menor a cero*
Presión Iny BN	*DiámetroEstrang*	*Menor a cero*
Presión Bajante	*DiámetroEstrang*	*Menor a cero*
Presión Separador	*DiámetroEstrang*	*Menor a cero*

Tabla 6.9. Continuación.

Parámetro X	Parámetro Y	R^2
Presión Salida	*DiámetroEstrang*	*Menor a cero*
Temp Separador	DiámetroEStrang	Menor a cero
QInyBN	DiámetroEStrang	Menor a cero
Qg Total (BN+Formación)	DiámetroEStrang	Menor a cero
QgForm	DiámetroEStrang	Menor a cero
RGA	DiámetroEStrang	Menor a cero
RGIL	DiámetroEStrang	Menor a cero
Presión Iny BN	PTP	Menor a cero
Presión Bajante	PTP	0.92
Presión Separador	PTP	Menor a cero
Presión Salida	PTP	Menor a cero
Temp Separador	PTP	Menor a cero
QInyBN	PTP	Menor a cero
Qg Total (BN+Formación)	PTP	Menor a cero
QgForm	PTP	Menor a cero
RGA	PTP	Menor a cero
RGIL	PTP	Menor a cero
Presión Separador	Presión Iny BN	Menor a cero
Presión Salida	Presión Iny BN	Menor a cero
Temp Separador	Presión Iny BN	Menor a cero
QInyBN	Presión Iny BN	Menor a cero
Qg Total (BN+Formación) Total	Presión Iny BN	0.881
QgForm	Presión Iny BN	0.887
RGA	Presión Iny BN	0.818
RGIL	Presión Iny BN	Menor a cero
Presión Separador	Pbajante	Menor a cero
Presión Salida	Pbajante	Menor a cero
Temp Separador	Pbajante	Menor a cero
QInyBN	Pbajante	Menor a cero
Qg Total (BN+Formación)	Pbajante	Menor a cero
QgForm	Pbajante	Menor a cero
RGA	Pbajante	Menor a cero

Tabla 6.9. Continuación.

Parámetro X	Parámetro Y	R^2
RGIL	Pbajante	Menor a cero
Presión Separador	Presión Iny BN	Menor a cero
Presión Salida	Presión Iny BN	Menor a cero
Temp Separador	Presión Iny BN	Menor a cero
QInyBN	Presión Iny BN	Menor a cero
Qg Total (BN+Formación) Total	Presión Iny BN	0.881
QgForm	Presión Iny BN	0.887
RGA	Presión Iny BN	0.818
RGIL	Presión Iny BN	Menor a cero
Presión Separador	Pbajante	Menor a cero
Presión Salida	Pbajante	Menor a cero
Temp Separador	Pbajante	Menor a cero
QInyBN	Pbajante	Menor a cero
Qg Total (BN+Formación)	Pbajante	Menor a cero
QgForm	Pbajante	Menor a cero
RGA	Pbajante	Menor a cero
RGIL	Pbajante	Menor a cero
Presión Salida	Presión Separador	Menor a cero
Temp Separador	Presión Separador	Menor a cero
QInyBN	Presión Separador	Menor a cero
Qg Total (BN+Formación)	Presión Separador	Menor a cero
QgForm	Presión Separador	Menor a cero
RGA	Presión Separador	Menor a cero
RGIL	Presión Separador	Menor a cero
Temp Separador	Presión Salida	Menor a cero
QInyBN	Presión Salida	Menor a cero
Qg Total (BN+Formación)	Presión Salida	Menor a cero
QgForm	Presión Salida	Menor a cero
RGA	Presión Salida	Menor a cero
RGIL	Presión Salida	Menor a cero
QInyBN	Temp Separador	Menor a cero
Qg Total (BN+Formación)	Temp Separador	Menor a cero

Tabla 6.9. Continuación

Parámetro X	Parámetro Y	R^2
QgForm	Temp Separador	Menor a cero
RGA	Temp Separador	Menor a cero
RGIL	Temp Separador	Menor a cero
Qg Total (BN+Formación)	QInyBN	Menor a cero
QgForm	QInyBN	Menor a cero
RGA	QInyBN	Menor a cero
RGIL	QInyBN	Menor a cero
QgForm	Qg Total (BN+Formación)	0.975
RGA	Qg Total (BN+Formación)	0.918
RGIL	Qg Total (BN+Formación)	Menor a cero
RGA	QgForm	0.936
RGIL	QgForm	Menor a cero
RGIL	RGA	Menor a cero

Continuando con el análisis de regresión se analizan las relaciones entre grupos de atributos. En el análisis se obtuvo que existe una correlación de un 68.4% entre el grupo de atributos formado por presión en el separador, presión en la salida, gasto de inyección de BN y gasto de aceite. Entre la presión en la tubería de producción y el gasto de gas se encontró una correlación de 98.41%, la cual se obtuvo con el módulo de análisis de regresión del software IDEA[1], y se muestra en la Figura 6.13.a). El gasto de gas a su vez tiene una correlación de 76.1% con el gasto de aceite (Tabla 6.9). Del lado izquierdo aparecen las variables seleccionadas y el resultado de las relaciones existentes entre ellas se muestran del lado derecho. Puede observarse en la última columna como están relacionadas estas variables. Esta correlación es físicamente correcta porque la presión en la tubería de producción está relacionada con la presión de fondo, la cual a su vez está relacionada con el gasto de aceite a través de la Ley de Darcy. Posteriormente, el mismo software nos permitió realizar un análisis estadístico avanzado, en el que se encontraron las relaciones existentes de cada uno de los parámetros contra todos los demás parámetros. Los resultados se muestran en la Figura 6.13.b), en donde puede observarse como las diferentes presiones están

correlacionadas, al igual que el gasto de inyección de BN. En esta figura se presenta una matriz en donde en las columnas se tienen todas las variables que participan en el sistema, mismas que aparecen en los renglones, de tal forma que el resultado mostrado en cada uno de los elementos de la matriz son las relaciones de cada parámetro contra cada parámetro. Puede observarse en la diagonal que los valores de 100 por ciento son obvios, al tratarse del la misma variable contra sí misma. Aquí deben rescatarse aquellas relaciones significativas, en donde el porcentaje sea mayor a un 50%.

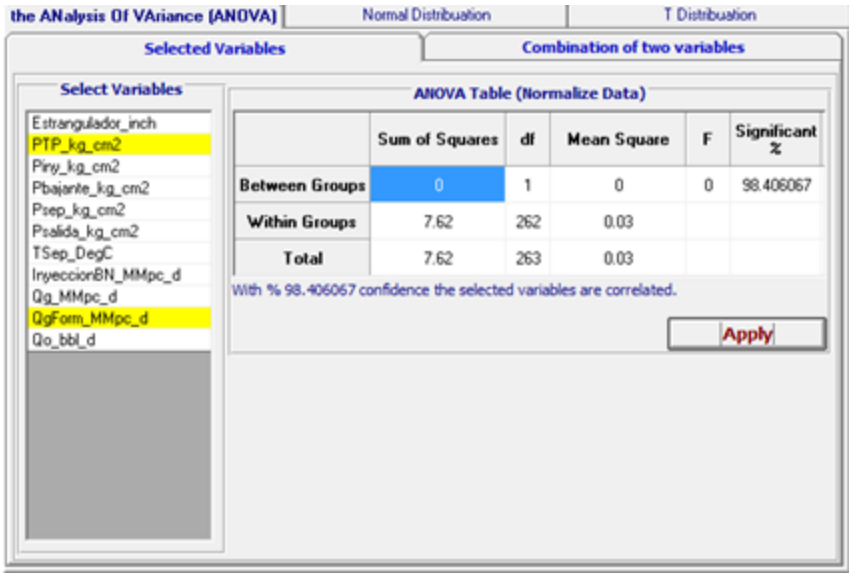


Figura 6.13. a) Correlación entre la presión en la tubería de producción y el gasto de gas de la formación.

	Estrangulador_inch	PTP_kg_cm2	Piny_kg_cm2	Pbajante_kg_cm2	Psep_kg_cm2	Psalida_kg_cm2	TSep_DegC	InyeccionBN_MMpc_d	Qg_MMpc_d	QgForm_MMpc_d	Qo_bbl_d
Piny_kg_cm2	42.8	0	100	0	0	0	0	0	0	0	0
Pbajante_kg_cm2	0	35.1	0	100	41.4	27.7	0	24	0	0	0
Psep_kg_cm2	0	90	0	41.4	100	78.1	0	67.6	0	0	0
Psalida_kg_cm2	0	88.1	0	27.7	78.1	100	0	87.5	0	0	0
TSep_DegC	0	0	0	0	0	0	100	0	0	0	0
InyeccionBN_MMpc_d	0	76.8	0	24	67.6	87.5	0	100	0	0	0
Qg_MMpc_d	0	0	0	0	0	0	0	0	100	7.4	0

Figura. 6.14 b) Resultados del análisis estadístico avanzado.

El resultado de esta etapa se puede considerar como un conjunto de datos listo para la aplicación del método de validación del dato de producción. Por otro lado, el análisis realizado permitió identificar los parámetros que más influyen en el comportamiento de la variable de salida (Q_o), presión en la tubería de producción, gasto de gas de la formación, gasto de inyección de gas de BN, y presión de inyección de BN.

6.2.II. Método de validación de datos

Una vez determinados los parámetros o atributos de mayor influencia para la salida del sistema, el paso siguiente fue aplicar la metodología y el algoritmo propuesto en el capítulo 5, sección 5.1.

La metodología se aplicó utilizando las herramientas que se presentan en la Tabla 6.10 Las tareas descritas son las incluidas en el algoritmo mencionado. En la primera columna de esta tabla tenemos una descripción corta de la tarea y en la segunda columna la herramienta utilizada. Es conveniente tener presente que para la aplicación de este método se utilizó nuevamente en el software *Intelligent Data Evaluation & Analysis*[1] (IDEA). Estas tareas pueden realizarse por otros productos de software diferentes a los indicados, como es el caso de los algoritmos Fuzzy C-Means y RNA que se incluyen en el software Matlab. Además, existen librerías de software libre que proveen esta misma funcionalidad, pero que requieren de conocimientos especializados de programación.

Tabla 6.10. Herramientas de apoyo al método de validación.

Tareas	Herramientas
6.2.II.1. Clasificación de los registros en tres clases	Software Intelligent Data Evaluation & Analysis[1], módulo de clasificación basado en algoritmo Fuzzy C-Means.
6.2.II.2. Entrenamiento de la Red Neuronal Artificial	Excel, macros, Software Intelligent Data Evaluation & Analysis[1], Módulo de RNA.
6.2.II.3. Proceso iterativo de validación	Excel (macro), programa desarrollado en lenguaje C#, Software Intelligent Data Evaluation & Analysis[1]

A continuación se describen con detalle cada una de las tareas mencionadas en la Tabla 6.10. Llevar a cabo cada una de ellas en la secuencia y orden propuesto nos permitirá cumplir nuestra meta: validar la calidad de los datos.

6.2.II.1. Clasificación de los registros en tres clases.

Esta inicia con la clasificación de los elementos del conjunto de datos preparado en la sección I de este capítulo, el cual se almacenó en la primera hoja, denominada "DatosAValidar", de un archivo en Excel. Esta hoja contiene los datos de origen, que son los valores reales que se requiere validar. Para llevar a cabo el proceso de clasificación, se seleccionan los atributos siguientes:

- Diámetro del estrangulador
- Presión en la tubería de producción
- Presión de inyección de BN
- Presión en la bajante
- Presión en el separador
- Presión a la salida
- Temperatura en el separador
- Gasto de inyección de BN
- Gasto de gas total (incluye gasto de inyección de BN mas gasto de formación)
- Gasto de gas de la formación
- Gasto de aceite

Se utilizó la herramienta de clasificación difusa que implementa el algoritmo Fuzzy C-Means descrito en el capítulo 4, que requiere el módulo de clasificación difusa del software IDEA[1]. Se agruparon los datos en tres clases diferentes, dándole a cada registro de datos un valor de pertenencia para cada una de las clases. Esto quiere decir que cada elemento del conjunto de datos pertenece en cierto grado a cada una de las tres clases. El grado de pertenencia a cada una de las clases debe ser un valor mayor o igual a 0 y menor o igual a 1. La suma de los tres valores debe ser 1. Si uno de los elementos tiene muy marcada su pertenencia a una de las clases, su grado de pertenencia a esa clase es 1 o casi 1.

La Figura 6.15 muestra como se incluyeron en el módulo de clasificación “difusa” del software IDEA[1] los parámetros de mayor influencia. En el lado izquierdo de la figura aparece el conjunto de atributos seleccionados para participar en la clasificación. Nótese que Q_o, el parámetro de salida de nuestro sistema, es incluido. En la sección central-inferior de la misma figura, se puede ver el valor asignado al parámetro o índice de difusividad del algoritmo Fuzzy C-Means, así como el valor asignado al número de clases.

Se clasificó cada uno de los registros en las tres clases mencionadas, obteniendo para cada uno de ellos el grado de pertenencia a cada una de las clases. La clase a la que “más” pertenece (clase dominante) es la que se considera en este análisis, así como su grado de pertenencia.

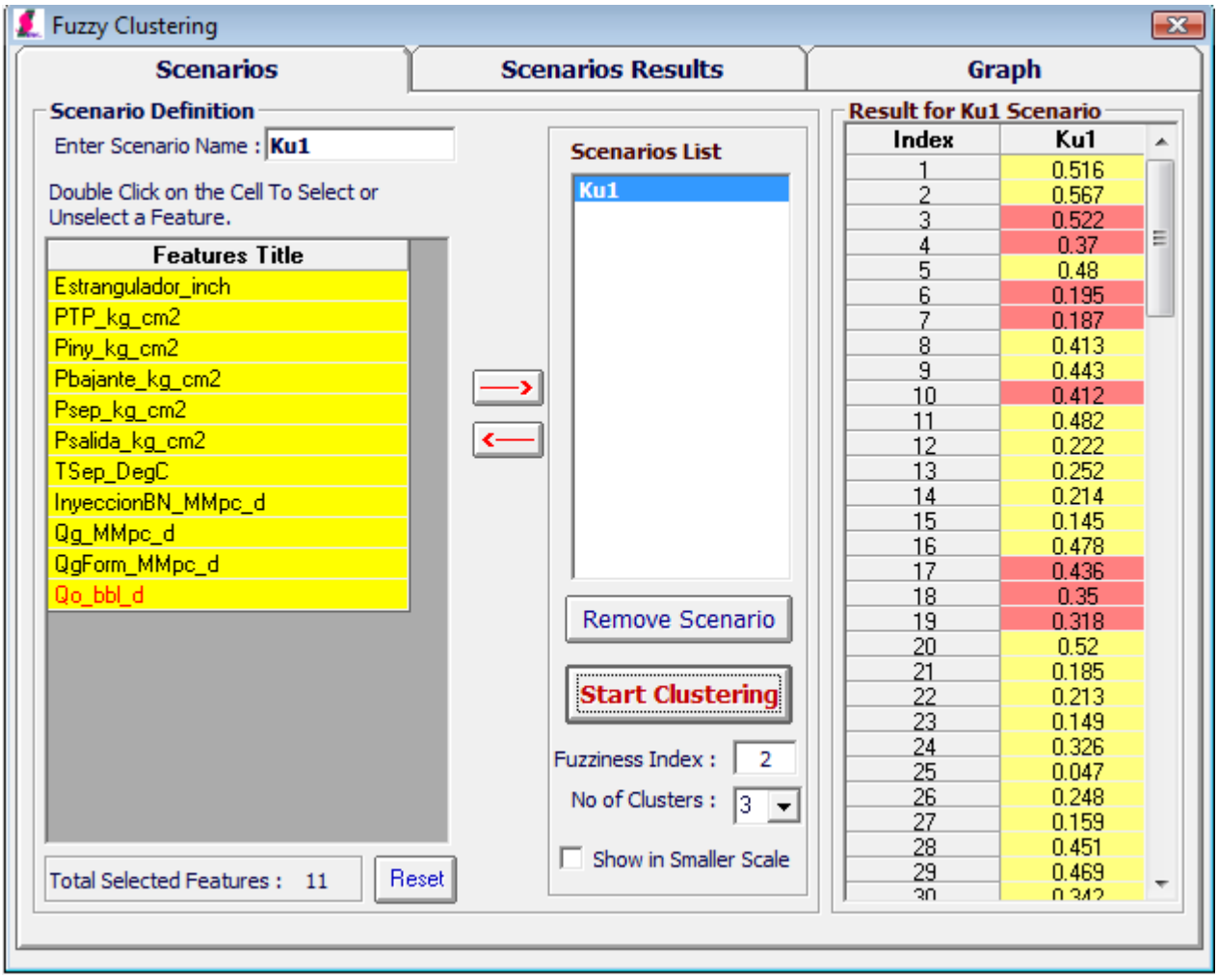


Figura 6.15. Escenario de clasificación difusa.

En la Figura 6.16 se presentan en la última columna (llamada Ku-1) los valores de la clase dominante obtenidos durante el proceso de clasificación difusa, mientras que en la Figura 6.17 se incluye el grado de pertenencia a esta clase dominante. Todos los resultados de la clasificación, clase dominante, grado de pertenencia y entropía de cada uno de los registros, se exportaron a Excel, de tal forma que pudieron utilizarse en la etapa siguiente de esta metodología.

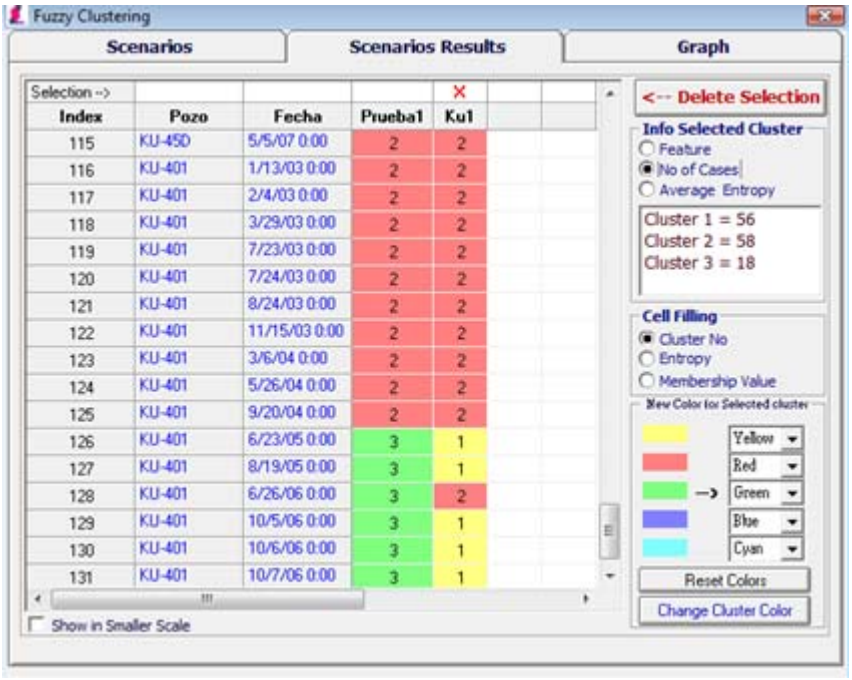


Figura 6.16. Clase dominante del conjunto de datos.

Figura 6.17. Entropía del conjunto de datos.

En la Tabla 6.11 se presenta un resumen de los resultados de la clasificación, que incluye las tres clases definidas, el número de registros en cada una de ellas y la entropía promedio de cada clase. Puede observarse que la clase con más elementos es la 2, además de ser la clase con menor valor de entropía. Si se considera la definición de entropía (la medida de la carencia de orden de un sistema), la clase 2 tienen el valor más cercano a cero, lo cual permite afirmar que es la clase con mayor orden.

Tabla 6.11. Resultado del proceso de clasificación difusa (Fuzzy C-Means)

Clase	Número de Registros	Entropía promedio
1	56	0.345
2	58	0.251
3	18	0.266
Todas las clases		**0.266**

6.2.II.2. Entrenamiento de la Red Neuronal Artificial.

El resultado de la tarea de clasificación se documentó en el archivo de Excel ya mencionado, en donde se agregó la hoja “Clasificados”, en la cual los parámetros de clasificación conjuntamente con los parámetros originales, se integraron para formar el

archivo que sirvió para entrenar la red neuronal artificial del siguiente punto de este método.

Como se mencionó, cada uno de los elementos del conjunto de datos está asociado a cada una de las clases por medio de un grado de pertenencia. La entropía de cada elemento se obtuvo dividiendo el valor mínimo de pertenencia entre el valor máximo.

Para integrar la información de clasificación, se agregaron tres columnas al conjunto de datos original: el valor de la clase dominante, el valor del grado de pertenencia a esta clase dominante, y el valor de la entropía.

Los parámetros clase, grado de pertenencia y entropía agregados, enriquecen el conjunto de datos original, porque el parámetro de salida (Q_o) está participando en forma indirecta. La estructura del nuevo conjunto de datos se presenta en la Tabla 6.12, la cual servirá para entrenar la red neuronal artificial.

Tabla 6.12. Registro de datos utilizado para entrenar la red neuronal.

Pozo	Fecha	DiaEstrang	PTP	Piny	Pbaj	Psep	Psal	TSep	InyeccionBN	Qo	QgT	Tbaj	QgF	Clase	Pertenencia	Entropia

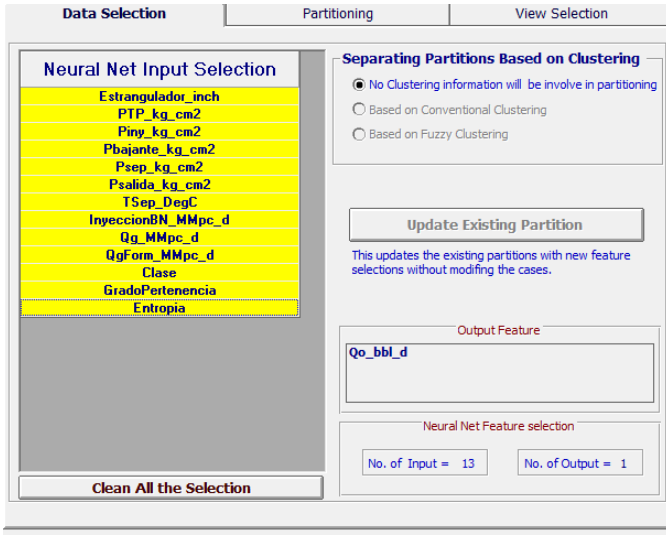
Para generar el nuevo modelo de predicción del gasto de salida, se entrenó una red neuronal artificial en donde participaron los parámetros de mayor influencia y los obtenidos durante el proceso de clasificación difusa. Es importante mencionar que este modelo es diferente al generado en la etapa de preparación de datos, ya que en aquella ocasión se utilizó para la estimación de datos faltantes. En este punto ya no nos preocupamos por los datos fuera de tendencia, ya que fueron eliminados en la etapa de preparación de datos.

A continuación se describen todas las actividades involucradas en el entrenamiento de la red neuronal artificial con el nuevo conjunto de datos.

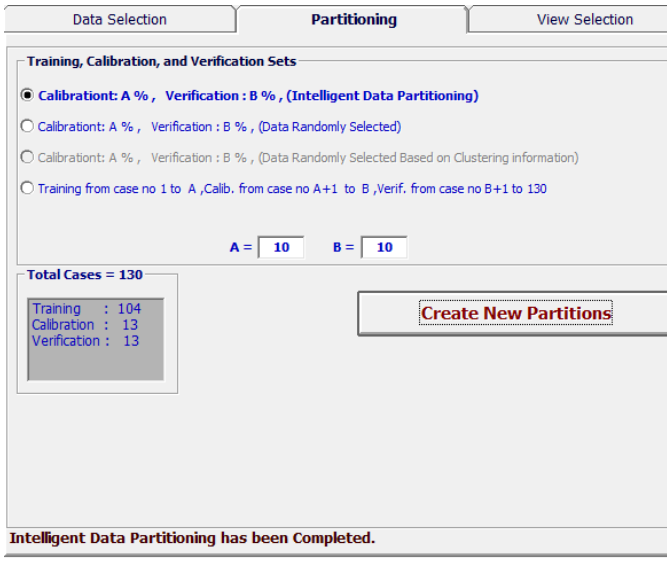
Preparación de la RNA

Esta actividad consistió en la separación de la totalidad de los datos en 3 subconjuntos: entrenamiento, calibración y verificación. Se utilizó un esquema de particionamiento inteligente que está disponible en el software IDEA[1]. Sin embargo, también se pueden definir otros esquemas de separación de los datos, como el aleatorio, con resultados

similares. En Figura 6.18 se muestra la interface del módulo de preparación de datos del software IDEA[1], en donde se observa cuales son los parámetros de entrada, la salida y en donde le indicamos cuales son las reglas para separar los datos. En este caso se le está indicando que el tipo de separación de datos sea inteligente. Una vez que se ejecuta este proceso, se visualiza como quedaron separados los datos: los datos marcados con T son para el entrenamiento, con V para la verificación y con C para la calibración. Se puede observar que el mayor número de elementos se usa para el entrenamiento.

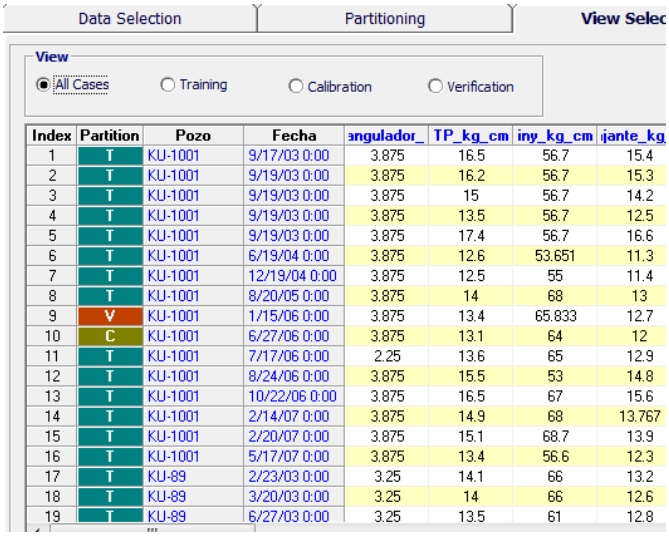


a) Selección de atributos



b) Partición inteligente de los datos.

c) Resultado de la partición



Index	Partition	Pozo	Fecha	anguladar_	TP_kg_cm	iny_kg_cm	iante_kg
1	T	KU-1001	9/17/03 0:00	3.875	16.5	56.7	15.4
2	T	KU-1001	9/19/03 0:00	3.875	16.2	56.7	15.3
3	T	KU-1001	9/19/03 0:00	3.875	15	56.7	14.2
4	T	KU-1001	9/19/03 0:00	3.875	13.5	56.7	12.5
5	T	KU-1001	9/19/03 0:00	3.875	17.4	56.7	16.6
6	T	KU-1001	6/19/04 0:00	3.875	12.6	53.651	11.3
7	T	KU-1001	12/19/04 0:00	3.875	12.5	55	11.4
8	T	KU-1001	8/20/05 0:00	3.875	14	68	13
9	V	KU-1001	1/15/06 0:00	3.875	13.4	65.833	12.7
10	C	KU-1001	6/27/06 0:00	3.875	13.1	64	12
11	T	KU-1001	7/17/06 0:00	2.25	13.6	65	12.9
12	T	KU-1001	8/24/06 0:00	3.875	15.5	53	14.8
13	T	KU-1001	10/22/06 0:00	3.875	16.5	67	15.6
14	T	KU-1001	2/14/07 0:00	3.875	14.9	68	13.767
15	T	KU-1001	2/20/07 0:00	3.875	15.1	68.7	13.9
16	T	KU-1001	5/17/07 0:00	3.875	13.4	56.6	12.3
17	T	KU-89	2/23/03 0:00	3.25	14.1	66	13.2
18	T	KU-89	3/20/03 0:00	3.25	14	66	12.6
19	T	KU-89	6/27/03 0:00	3.25	13.5	61	12.8

Figura 6.18. Preparación de la RNA en el software IDEA[1]

Entrenamiento de la RNA

Una vez separados los datos, se configuraron los parámetros de aprendizaje de la RNA. La RNA a entrenar tiene tres capas: una de entrada, una oculta y una de salida. El algoritmo de aprendizaje utilizado fue el de retro-propagación del error. Estos conceptos se describieron en el capítulo 4. La RNA se entrenó empleando el módulo de redes neuronales con retro-propagación del software IDEA[1], iniciando con la configuración predeterminada por el software IDEA[1]; se evaluaron diferentes escenarios para encontrar la configuración óptima, los cuales se presentan en la Tabla 6.13. Se puede observar que existe una configuración independiente para las interfaces entre capas. En las dos primeras columnas del lado izquierdo se muestran los valores de configuración para la interface entre las capas entrada-oculta. En las siguientes dos columnas los valores para la interface entre las capas oculta- capa de salida. R^2 es el indicador utilizado para encontrar el mejor escenario. El escenario sombreado fue el que logró una R^2 más cercana al valor de 1.

Tabla 6.13. Diferentes escenarios evaluados para el entrenamiento de la RNA.

Capa entrada – capa oculta		Capa oculta – capa salida		Función Activación	Núm.Iteraciones	Número Neuronas	Valor inicial para generador de número aleatorios	R (calibración)	R^2 (calibración)
Momento	Razón de aprendizaje	Momento	Razón de aprendizaje						
0.8	0.1	0.8	0	Logística	39	25	1	0.9758	0.9389
0.8	0.1	0.8	0	Logística	39	30	1	0.9700	0.8830
0.8	0.01	0.8	0	Logística	530	25	1	0.9754	0.9491
0.8	0.05	0.8	0	Logística	98	25	1	0.9757	0.9460
0.8	0.05	0.8	0	Logística	593	17	1	0.9764	0.9465
0.8	0.05	0.8	0.2	Logística	1000	30	1	0.9685	0.9308

En la Figura 6.19 puede observarse el comportamiento de R^2 para los diferentes escenarios listados en la Tabla 6.13.

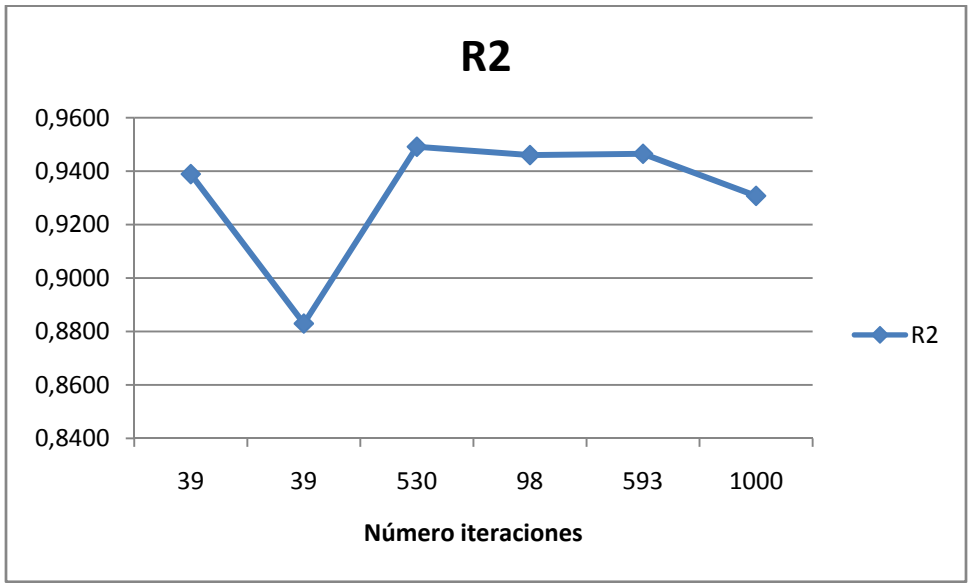


Figura 6.19. Comportamiento de R^2 para los diferentes escenarios evaluados.

Para configurar la red se indicaron los valores óptimos de momento, razón de aprendizaje y decaimiento del peso obtenidos, para cada una de las conexiones sinápticas entre las capas (Entrada-Oculta y Oculta-Salida). Estas conexiones sinápticas pueden observarse en la topología de la red presentada en la Figura 6.20.

También se indicó el tipo de función de activación, ya sea logística o sigmoidal, descritas en el Capítulo 4.

Los valores predeterminados por el software se calculan con base en el número de entradas configuradas en la etapa de preparación de la red neuronal. Empleando esta información y el número de registros del conjunto de datos, se calcula el número de neuronas en la capa oculta. Los números aleatorios se usan para inicializar los pesos en la red neuronal antes del entrenamiento. Estos valores de configuración se modificaron conforme a los resultados mostrados en la Tabla 6.13. El diseño de la red neuronal artificial en el software IDEA[1] se muestra en la Figura 6.18.

En la Tabla 6.13 puede observarse que la configuración óptima corresponde a los valores de razón de aprendizaje de 0.01, momento de 0.8 entre la capa de entrada y la capa oculta. Para la interface capa oculta – capa de salida, un momento de 0.8 y la razón de aprendizaje fue cero. Este valor cero para la razón de aprendizaje indica que los pesos de las conexiones capa oculta – capa salida no se alteraron durante el proceso de entrenamiento. El proceso de aprendizaje realmente ocurrió entre las conexiones capa entrada –capa oculta, y fue lo que determinó el resultado obtenido. El número de iteraciones óptimo fue de 530.

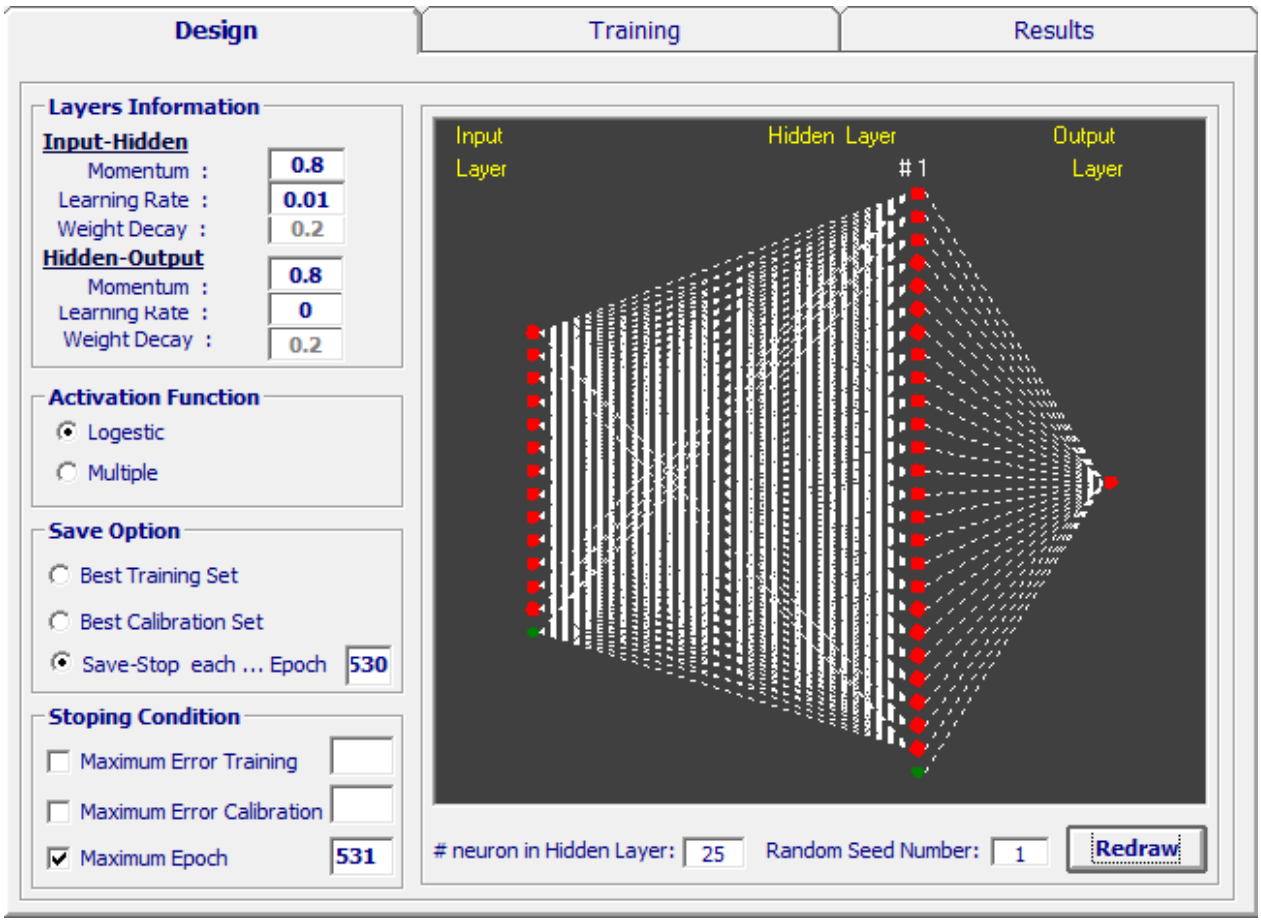


Figura 6.20. Diseño de la RNA con los parámetros de configuración óptima utilizando software IDEA[1].

Como se mencionó en el capítulo 4, el modelo de retro-propagación busca minimizar el error (diferencia obtenida entre la salida estimada y la deseada). En la Figura 6.21 puede observarse como durante el entrenamiento de la RNA, el error disminuye conforme aumenta el número de iteraciones. En ambas gráficas, el eje de las *x* corresponde al número de iteraciones, mientras que el eje de las *y* corresponde al valor del error. En la primera gráfica se observa el error obtenido en el conjunto de datos de calibración, y en la segunda, el error calculado para el conjunto de datos de entrenamiento.

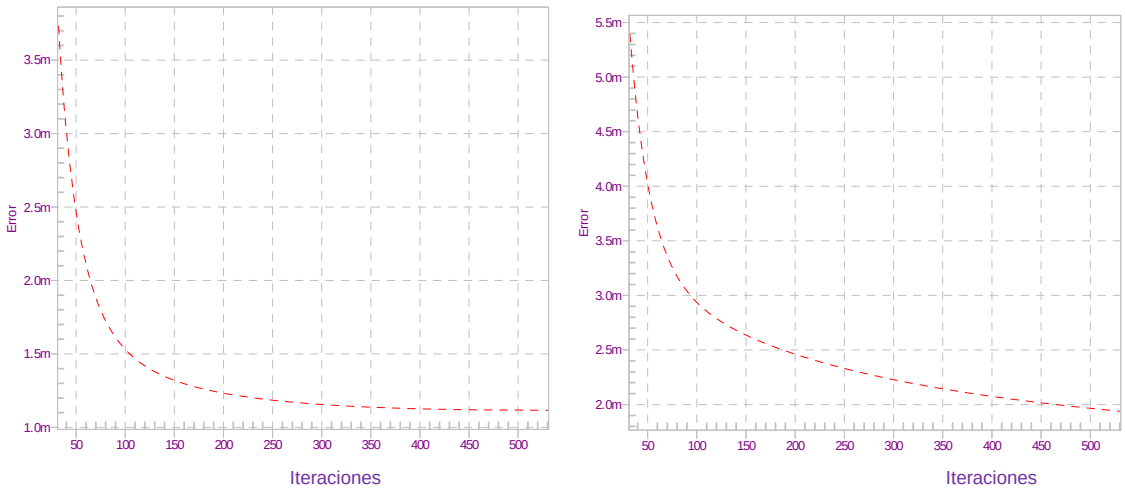


Figura 6.21. Comportamiento del error en los datos de calibración y en los de entrenamiento

Finalmente, en la Figura 6.22 se muestran los resultados de los tres subconjuntos de datos: entrenamiento, calibración, verificación y la suma de estos, así como el valor de R^2 para cada uno de ellos. En estas cuatro gráficas se puede observar el comportamiento de los valores reales (marcados con triángulos) con respecto a los valores de predicción de la RNA (marcados con círculos).

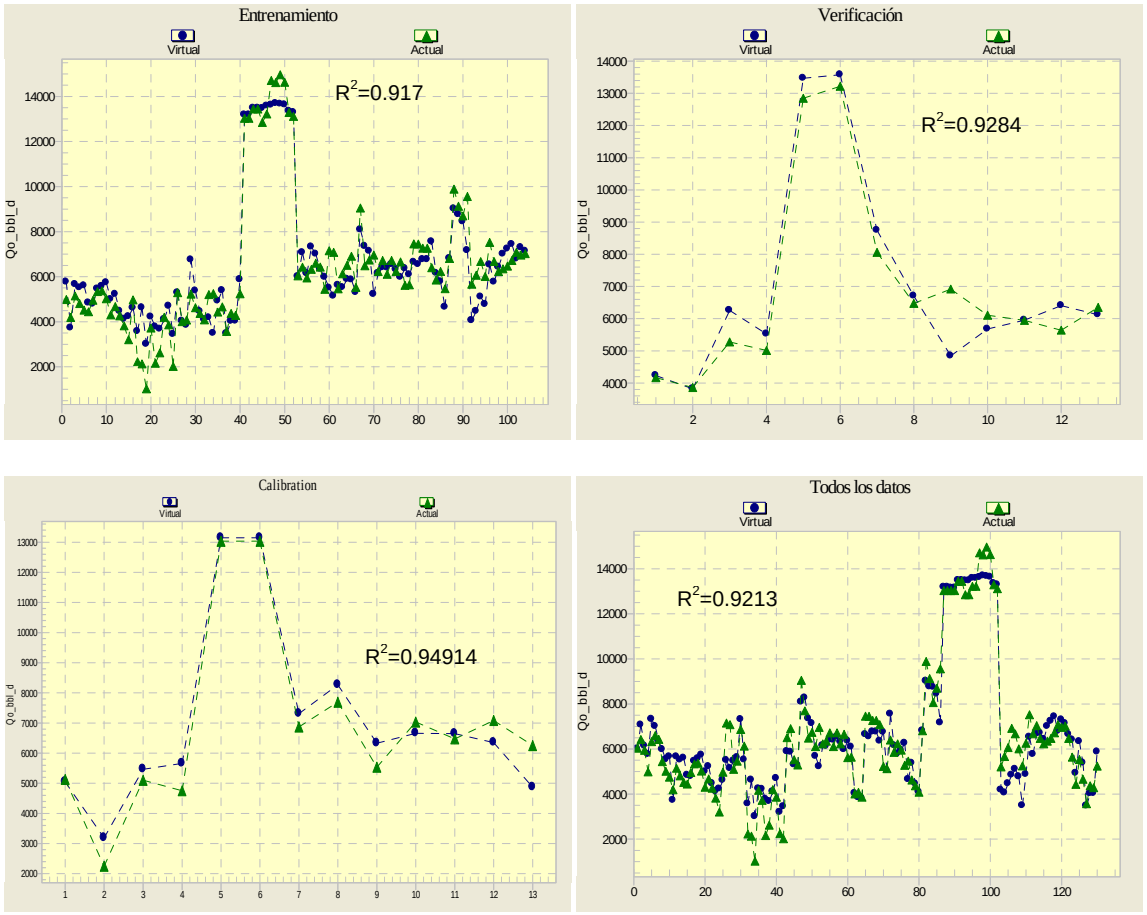


Figura 6.22. Resultado del proceso de entrenamiento de la RNA. La gráfica muestra el comportamiento del gasto Q_o real contra el valor de predicción de la RNA.

Este modelo es el que representa el sistema como un todo, ya que la salida del sistema, el gasto de aceite, participa en forma indirecta en el mismo. El modelo se guarda para usarse posteriormente.

Con el fin de validar si el modelo de la RNA desarrollado representaba nuestro sistema físico, se realizó un análisis de su comportamiento. En la Figura 6.23 se observa como el gasto de aceite se incrementa cuando el gasto de gas de formación también aumenta. Puede notarse en la Figura 6.24 que no existe correlación entre el gasto de inyección de BN y el gasto de aceite. Esto también se observa en la Figura 6.25 para el caso del pozo Ku-1001.

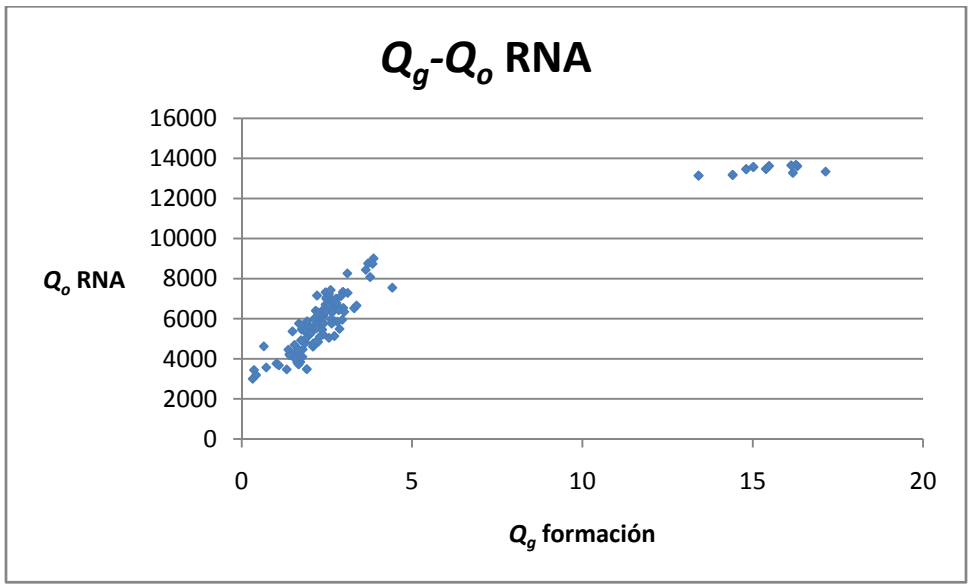


Figura 6.23. Relación del gasto de aceite estimado por el modelo de RNA y el gasto de gas de la formación

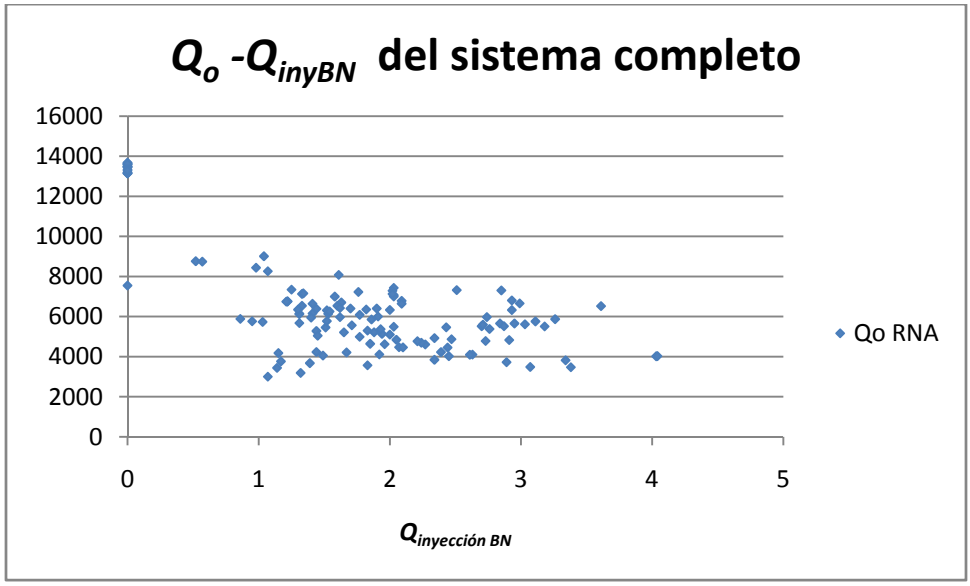


Figura 6.24. Comportamiento del gasto Q_o con respecto al gasto de inyección de BN, considerando todos los pozos del conjunto de datos.

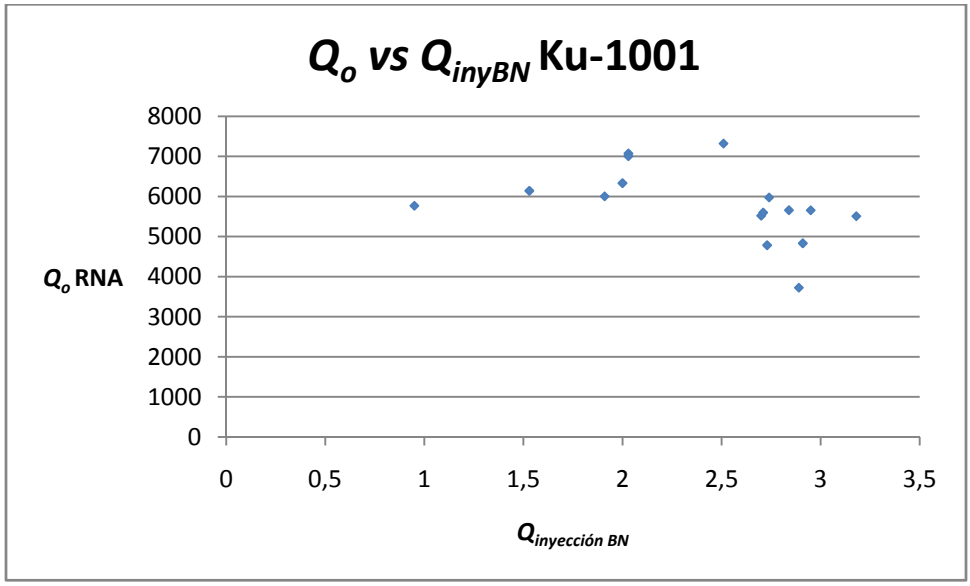


Figura 6.25. Comportamiento del gasto Q_o de la RNA con respecto al gasto de inyección de BN para el pozo KU-1001

El resultado de la segunda etapa de este método fue la red neuronal artificial entrenada, la cual quedó disponible para aplicarse en un proceso iterativo que ayudó a determinar la calidad de los elementos del conjunto de datos. Los valores de predicción de la RNA se guardaron en una columna adicional en la hoja “Clasificados”, porque fueron valores que se emplearon en el análisis de calidad de los datos.

6.2.II.3. Proceso iterativo para validación de datos

El paso siguiente fue recorrer el rango de salida, tomando los valores reales del gasto de aceite de todos los elementos del conjunto de datos.

- Gasto de aceite mínimo = 1,025 bbld
- Gasto de aceite máximo = 14,965 bbld

Se consideraron los 130 valores del gasto de aceite del rango de salida original, los cuales se denominan valores de salida supuestos. Posteriormente, para cada uno de los elementos del conjunto de datos original, se ejecutó el proceso iterativo siguiente. A continuación utilizaremos la variable *k* para referirnos al índice del elemento que se está validando.

3.1. Se seleccionó el elemento *k* del conjunto de datos y se recorrió el rango de salida del sistema. Esto se implementó en una hoja de Excel en donde se repitió 130 veces el elemento *k* y se sustituyó el valor de salida real por los 130 valores supuestos. El resultado fue el conjunto de datos del elemento *k*, que se denomina Conjunto *k*.

3.2. Se clasificó el Conjunto *k* siguiendo el mismo procedimiento documentado en el punto 6.2.II.1. Se obtuvo la información de clasificación (clase dominante, grado de pertenencia a la clase dominante y entropía) para cada uno de los elementos de este conjunto. Se agregaron tres columnas al Conjunto *k* con la información obtenida del proceso de clasificación.

3.3. Se aplicó el modelo de la red neuronal artificial entrenada en el punto 6.2.II.2 a los datos del Conjunto *k* clasificados. Se agregó una columna al Conjunto *k* que incluyó los valores de predicción de la RNA.

3.4. Se calculó el error para cada uno de los elementos del Conjunto *k*, donde:

$$Error = |SalidaRedNeuronal - SalidaSupuesta|$$

La variable de *SalidaRedNeuronal* representa el valor de predicción de la RNA y *SalidaSupuesta* el valor de salida supuesto.

3.5. Se generaron tres curvas con los datos del Conjunto *k*, donde cada punto de las curvas correspondió a un valor de la salida supuesta (eje X) :

Curva 1: Identifica la clase a la cual pertenece el punto.

Curva 2: Identifica la entropía del punto dentro de la clase a la que pertenece.

Curva 3: Identifica la diferencia entre el valor real y el valor de predicción de la red neuronal.

3.6. Se graficaron como dos líneas verticales los valores de salida del elemento *k* originales: gasto de aceite real (Q_{oReal}) y gasto de aceite estimado por la RNA (Q_{oRNA}) en el punto 6.2.II.2.

3.7. Se presentaron las tres curvas en una gráfica conjuntamente con las líneas verticales que representaban las salidas del sistema.

3.8. Se identificó la posición donde el error era cercano al valor cero y la entropía era mínima para una clase determinada (existe un punto de entropía mínima por cada clase).

3.9. Se identificó la clase dominante con base en el número de valores Q_o supuestos que pertenecían a cada clase.

Se repitió el proceso a partir del paso 3.1 para cada uno de los registros del conjunto de datos original.

La Tabla 6.14 muestra los atributos considerados durante los procesos de clasificación difusa y entrenamiento de la RNA descritos en los puntos 6.2.II.1 y 6.2.II.2.

Tabla 6.14. Atributos considerados en la clasificación difusa y en entrenamiento de la RNA.

Descripción del parámetro	**Nombre en la BD**	**Unidades**	**Rol**
Nombre del Pozo	Pozo		Id
Fecha de toma de medición	Fecha	dd/mm/aaaa	Id
Diámetro de Estrangulador	Estrangulador_inch	pulgadas	Entrada
Presión en la tubería de producción	PTP_kg_cm2	kg/cm^2	Entrada
Presión de inyección de bombeo neumática	Piny_kg_cm2		Entrada
Presión en la bajante	Presión ePbajante_kg_cm2	kg/cm^2	Entrada
Presión en el separador	Psep_kg_cm2	kg/cm^2	Entrada
Presión a la salida	Psalida_kg_cm2	kg/cm^2	Entrada
Temperatura en Separador	TSep_DegC	°C	Entrada
Gasto de inyección de BN	InyeccionBN_MMpc_d	MMpc/d	Entrada
Gasto de aceite	SeepQo_bbl_d	Bbl/d	Salida, rango de salida total
Gasto de gas total	Qg_MMpc_d	MMpc/d	Entrada
Gasto de gas de la formación	QgForm_MMpc_d	MMpc/d	Entrada

En la Figura 6.26 se puede observar la selección de atributos en el módulo de clasificación Fuzzy C-Means del software IDEA[1]. En la primera columna de esta figura, el software muestra los atributos o parámetros disponibles en el archivo de datos de entrada, y permite indicar en la segunda columna su rol dentro del sistema. Aquí se

puede indicar si es un parámetro identificador de cada elemento, si es de entrada, de salida o si no participa en el proceso.

Variables	Attribute	Min	Max	Mean	Stand. DV	No. Data Point
Pozo	ID	Not Numeric	Not Numeric	Not Numeric	Not Numeric	Not Nume
Fecha	ID	Date	Date	Date	Date	Date
Estrangulador_inch	Input	3.875	3.875	3.875	0	130
PTP_kg_cm2	Input	12.6	12.6	12.6	0	130
Piny_kg_cm2	Input	53.6505	53.6505	53.651	0	130
Pbajante_kg_cm2	Input	11.3	11.3	11.3	0	130
Psep_kg_cm2	Input	10	10	10	0	130
Psalida_kg_cm2	Input	9.2	9.2	9.2	0	130
TSep_DegC	Input	83	83	83	0	130
InyeccionBN_MMpc_d	Input	2.03	2.03	2.03	0	130
Real_Qo_bbl_d	Not Used	6596.03	6596.03	6596.03	0	130
Qg_MMpc_d	Input	4.52	4.52	4.52	0	130
QgForm_MMpc_d	Input	2.48	2.48	2.48	0	130
Clase Ori	Not Used	1	1	1	0	130
MSV Ori	Not Used	0.811	0.811	0.811	0	130
Entropía Ori	Not Used	0.195	0.195	0.195	0	130
SweepQo_bbl_d	Output	1025	14965	6624.255	2988.321	130
Qo_RNAClasificacionOriginal	Not Used	7006.726	7006.726	7006.726	0	130
Clase	Not Used	Not Numeric	Not Numeric	Not Numeric	Not Numeric	Not Nume
Pertenencia	Not Used	Not Numeric	Not Numeric	Not Numeric	Not Numeric	Not Nume
Entropia	Not Used	Not Numeric	Not Numeric	Not Numeric	Not Numeric	Not Nume

Figura 6.26. Asignación de atributos en el software IDEA[1].

En la Figura 6.27 se observan los parámetros que participaron en el proceso de clasificación, así como el número de clases y el índice de difusividad. Para cada elemento *k* se tuvo que repetir el procedimiento descrito y crear un proyecto diferente en el software IDEA[1].

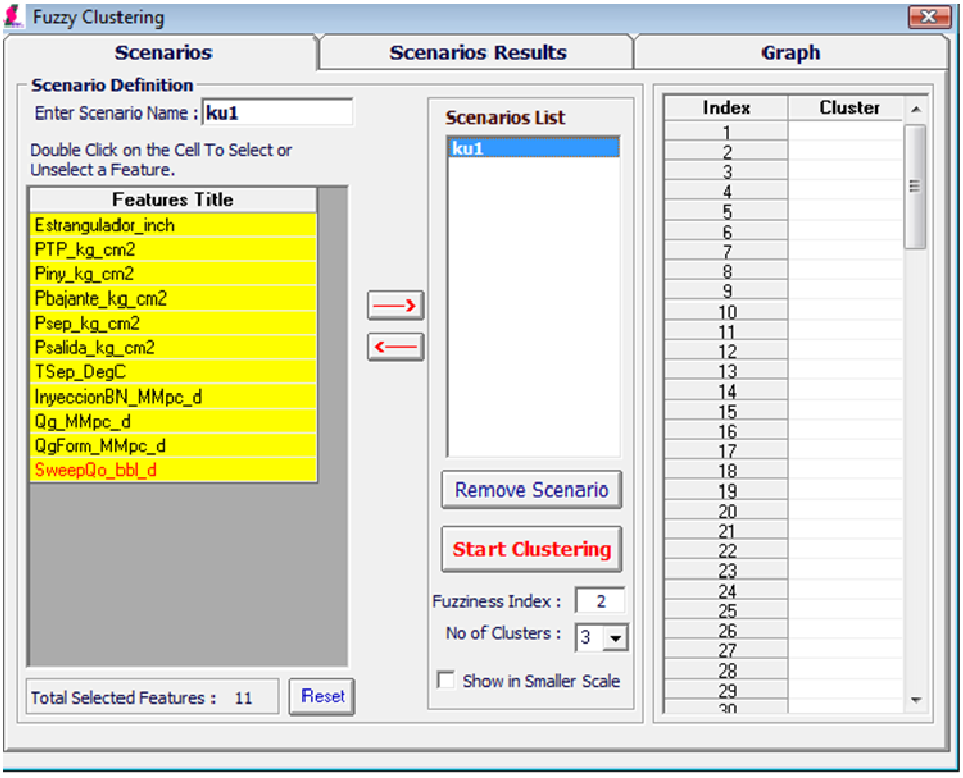


Figura 6.27. Configuración para la clasificación difusa con el software IDEA[1].

La Figura 6.28 ilustra cómo los valores de clasificación son exportados.

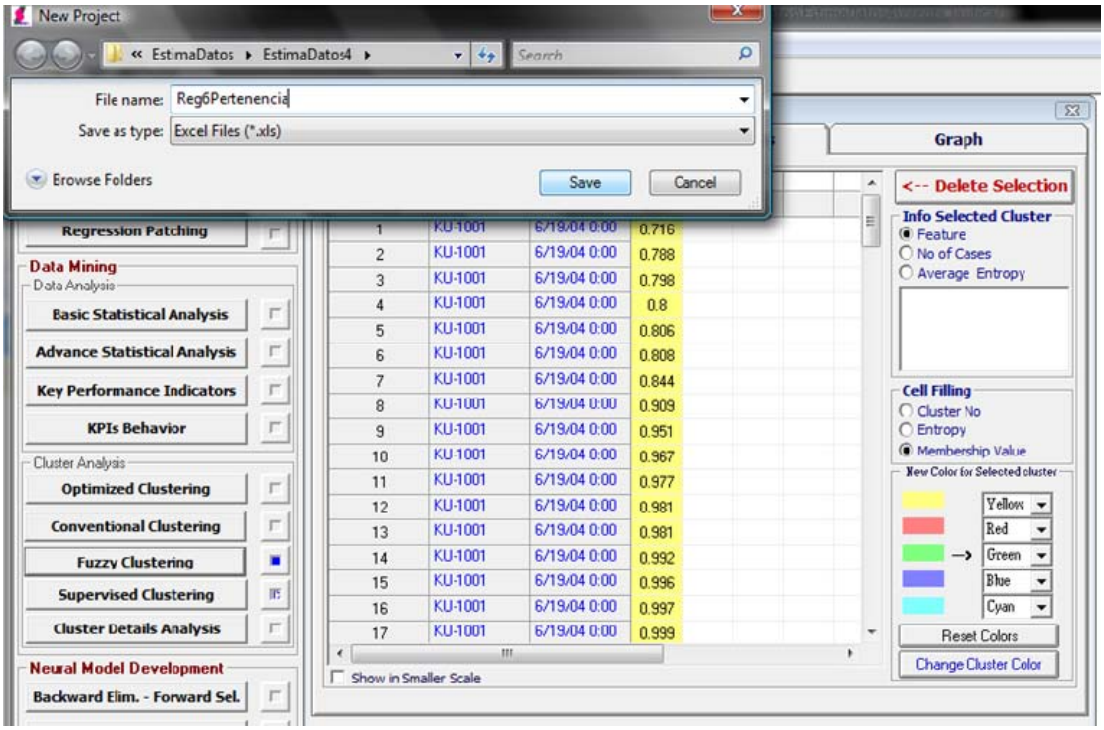
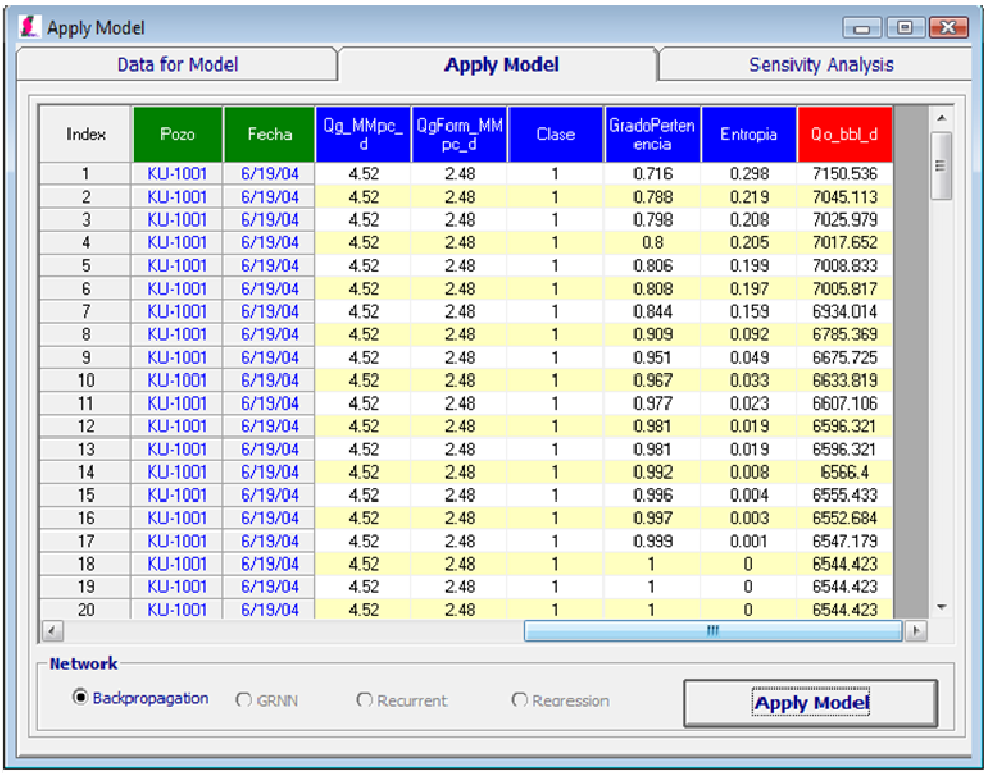


Figura 6.28. Exportación de la información de clasificación.

Los valores de clasificación se agregaron al Conjunto *k,* que fueron los datos empleados al aplicar el modelo de la RNA. En la Figura 6.29 se observa cómo estos atributos constituyen la entrada de la RNA entrenada en el punto 6.2.II.2, en donde se incluyen la información de clase, entropía y grado de pertenencia, y cómo participa en forma indirecta la salida del sistema, el gasto de aceite.



Index	Pozo	Fecha	Qg_MMpc_d	QgForm_MMpc_d	Clase	GradoPertenencia	Entropia	Qo_bbl_d
1	KU-1001	6/19/04	4.52	2.48	1	0.716	0.298	7150.536
2	KU-1001	6/19/04	4.52	2.48	1	0.788	0.219	7045.113
3	KU-1001	6/19/04	4.52	2.48	1	0.798	0.208	7025.979
4	KU-1001	6/19/04	4.52	2.48	1	0.8	0.205	7017.652
5	KU-1001	6/19/04	4.52	2.48	1	0.806	0.199	7008.833
6	KU-1001	6/19/04	4.52	2.48	1	0.808	0.197	7005.817
7	KU-1001	6/19/04	4.52	2.48	1	0.844	0.159	6934.014
8	KU-1001	6/19/04	4.52	2.48	1	0.909	0.092	6785.369
9	KU-1001	6/19/04	4.52	2.48	1	0.951	0.049	6675.725
10	KU-1001	6/19/04	4.52	2.48	1	0.967	0.033	6633.819
11	KU-1001	6/19/04	4.52	2.48	1	0.977	0.023	6607.106
12	KU-1001	6/19/04	4.52	2.48	1	0.981	0.019	6596.321
13	KU-1001	6/19/04	4.52	2.48	1	0.981	0.019	6596.321
14	KU-1001	6/19/04	4.52	2.48	1	0.992	0.008	6566.4
15	KU-1001	6/19/04	4.52	2.48	1	0.996	0.004	6555.433
16	KU-1001	6/19/04	4.52	2.48	1	0.997	0.003	6552.684
17	KU-1001	6/19/04	4.52	2.48	1	0.999	0.001	6547.179
18	KU-1001	6/19/04	4.52	2.48	1	1	0	6544.423
19	KU-1001	6/19/04	4.52	2.48	1	1	0	6544.423
20	KU-1001	6/19/04	4.52	2.48	1	1	0	6544.423

Figura 6.29. Resultado de aplicar el modelo de la RNA.

Los valores de predicción de RNA se agregaron al Conjunto *k*. Se calculó el error como la diferencia entre el valor de salida de la RNA y el valor de salida supuesto. Finalmente, se agregó una columna con el error calculado. En este punto ya se contaba con toda la información necesaria para calificar el elemento *k* de acuerdo a los criterios definidos en el Capítulo 5.

Para determinar la calidad del elemento *k* se graficaron las tres curvas (clase, entropía y error) y las líneas verticales (Q_{oReal} y Q_{oRNA}) conjuntamente con los datos del Conjunto *k* (puntos 3.5 al 3.9). La curva de clase se representa como una línea continua, y solo puede tener los valores 1,2, o 3. La curva de entropía es la línea discontinua, y puede adquirir valores entre 0 y 1. Finalmente, la curva de error es la línea punteada y puede tener valores mayores o iguales a cero. La línea vertical que representa la salida real (Q_{oReal}) se marcó con una cruz y la que representa el valor de salida estimado por la red neuronal (Q_{oRNA}) con un asterisco.

Finalmente, se aplicaron los criterios de calidad de los datos. Un registro se calificó como “bueno” cuando los valores de la curva de error y los de la curva de entropía fueron cercanos a cero en el mismo valor de la salida supuesta (eje *X*); y además, este valor de salida pertenecía a la clase dominante. La Figura 6.30 presenta un ejemplo de un registro calificado como bueno, en donde puede observarse que el valor Q_o supuesto que corresponde al error mínimo (marcado con un círculo) también corresponde a un valor de entropía cercano a cero y este valor de Q_o supuesto pertenece a la clase 2, que fue la clase dominante. El número de valores de la clase 2 fue ligeramente mayor que el de las otras dos clases. Además, la línea que representa el valor de salida real (Q_{oReal}), la línea que representa la salida estimada por la RNA (Q_{oRNA}), y el valor cero en la curva de error, son cercanos. Puede observarse como las líneas que representan las salidas (Q_{oReal} = 8062 y Q_{oRNA} = 7993) casi se traslapan. Los puntos marcados con una cruz en la curva de entropía corresponden al valor de entropía mínima de cada una de las clases. El punto marcado con un círculo en la curva de error corresponde al valor de error mínimo.

Otra manera de expresar el criterio para calificar un registro como “Bueno” sería:

SI *Entropía* ($Q_{oSupuestoErrorMínimo}$) < *CriterioEntropía*

Y

Clase ($Q_{oSupuestoErrorMínimo}$) = *ClaseDominante*

Y

$|Q_{oReal} - Q_{oRNA}|$ < *CriterioError*

ENTONCES

Registro “Bueno”

Donde:

$Q_{oSupuestoErrorMínimo}$ es el valor de Q_o supuesto (eje *X*) para el valor mínimo de la curva de error,

ClaseDominante es la clase con mayor número de elementos (determinado por el número de valores Q_o supuestos que pertenecen a cada clase),

CriterioEntropía es un valor cercano a cero definido por el analista, y

CriterioError también es un valor cercano a cero definido por el analista, el cual dependerá del comportamiento de la curva de error.

En esta aplicación del método de validación, el valor del criterio de entropía fue igual a 0.20; mientras que el criterio de error fue de 356, 5% del rango de valores de la curva de error (valor mínimo = 156, valor máximo error= 7278).

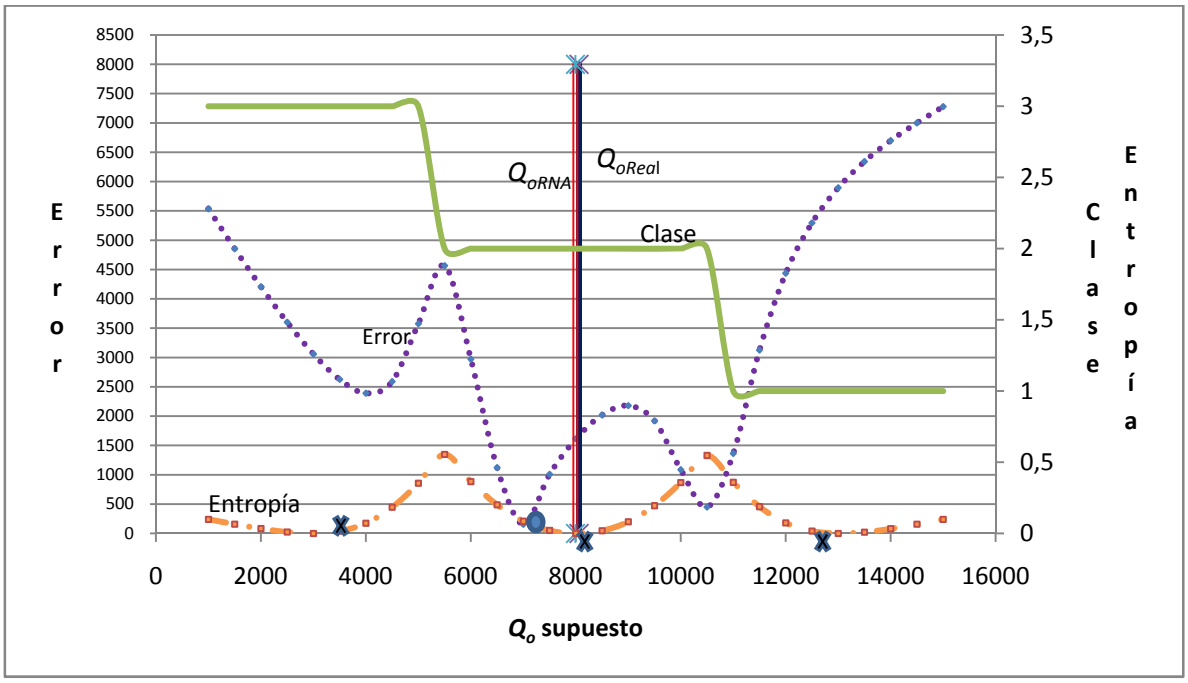


Figura 6.30. Comportamiento de un registro calificado como bueno.

Un registro se considera "ligeramente contaminado" cuando uno de los valores de salida, el valor de salida real (línea Q_{oReal}) o el valor de la salida de la red neuronal (línea Q_{oRNA}), es cercano al valor cero en la curva de error, y además cualquiera de estos dos valores de salida es cercano al valor mínimo de la curva de entropía. Esto expresado en otros términos quedaría:

SI

[

$Error\ (Q_{oReal}) < CriterioError$

O

$Error\ (Q_{oRNA}) < CriterioError$

]

Y

[

Entropía (Q_{oReal}) < CriterioEntropía

O

Entropía (Q_{oRNA}) < CriterioEntropía

]

Y

$|Q_{oReal} - Q_{oRNA}|$ < *CriterioError*

ENTONCES

Registro “Ligeramente Contaminado”.

Como en el caso de Registro Bueno, el criterio de error y el criterio de entropía son valores cercanos a cero y pueden definirse por el analista responsable de aplicar el método. Esto da un grado de subjetividad al método de análisis, ya que la calidad de los datos evaluados dependerá de este valor. En este ejercicio el criterio de entropía fue de 0.20, mientras que el error fue de 924, calculado con base en el rango de valores de la curva de error (10%).

Puede observarse en la Figura 6.31 que en el caso de un registro ligeramente contaminado, puede haber una diferencia entre la línea vertical que representa el valor de salida real y la línea vertical que representa la predicción de la RNA (Q_{oReal} = 5011 y Q_{oRNA} = 5522). La línea vertical que representa Q_{oReal} casi cruza con el valor de error mínimo y además cruza la línea de entropía en un valor cercano a cero (valor de entropía para Q_{oReal} = 0.199).

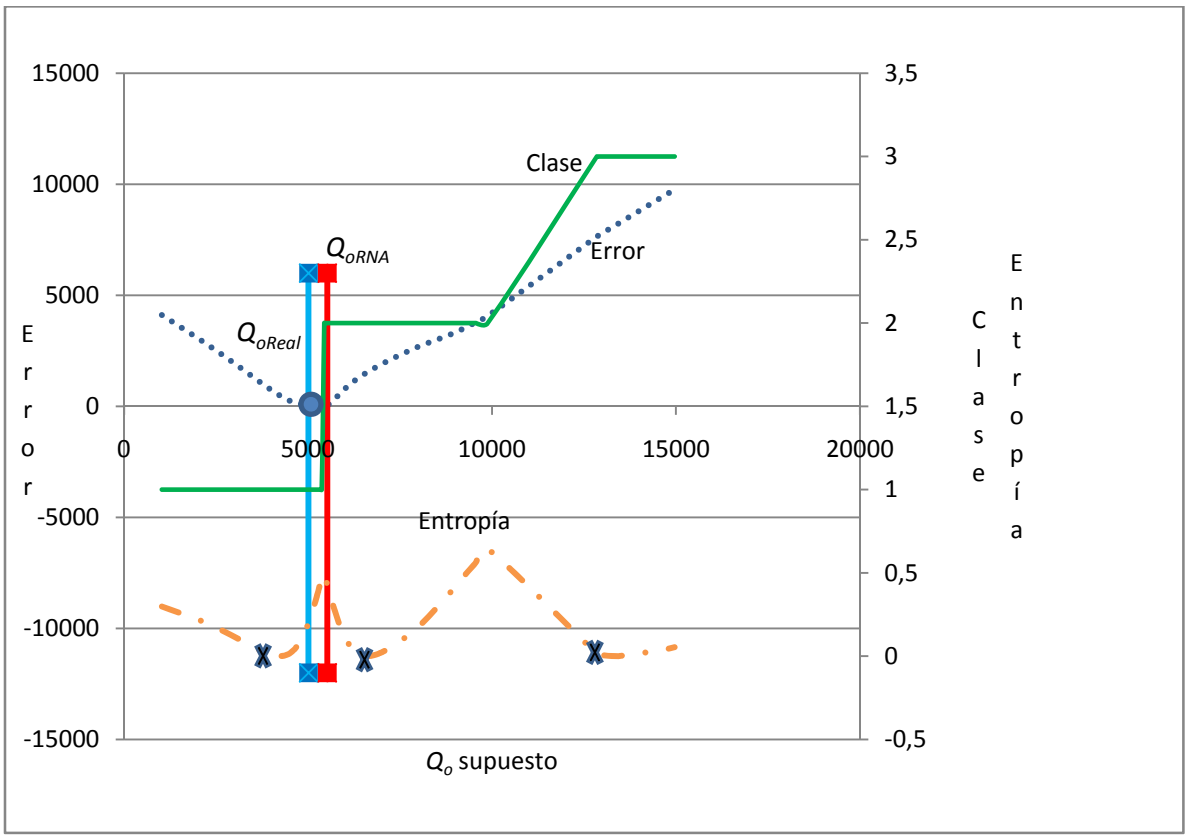


Figura 6.31. Comportamiento de un registro calificado como ligeramente contaminado.

En la Figura 6.32 se presenta un registro calificado como “malo”, ya que no existe correlación entre la salida original, la salida de la red neuronal y la curva de error. Además, existe una diferencia notable entre las dos líneas verticales que representan las salidas (Q_{oReal} y Q_{oRNA}), y no hay correlación de ambas líneas con la curva de entropía.

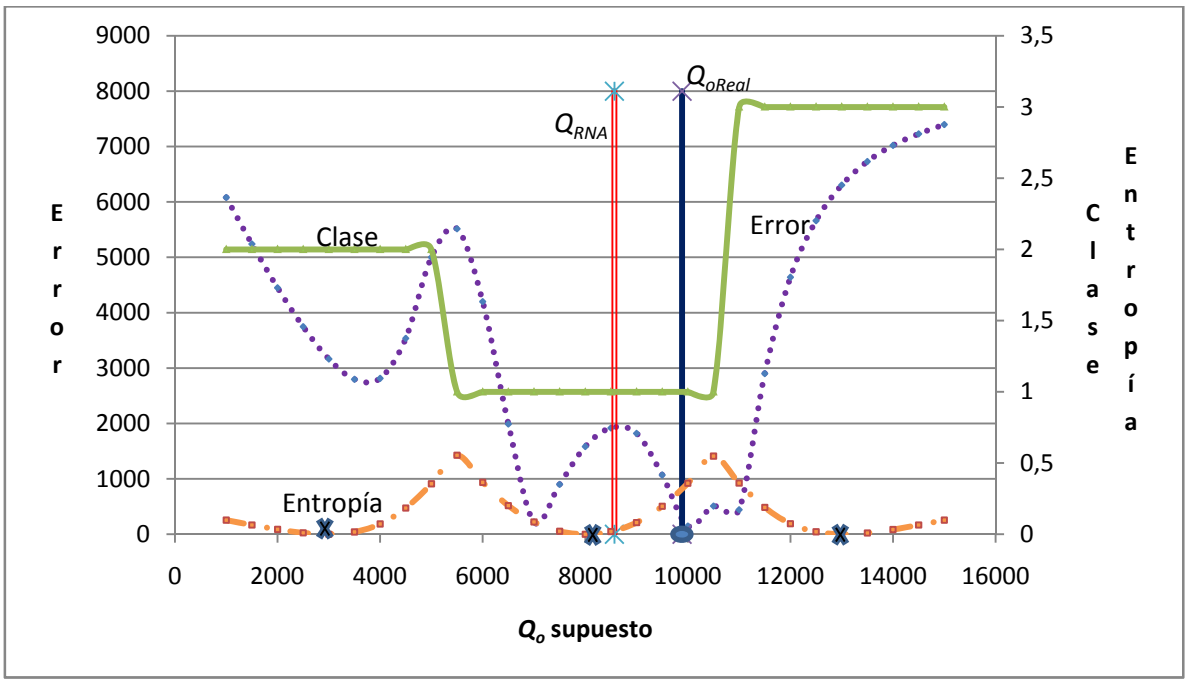


Figura 6.32. Comportamiento de un registro calificado como malo.

Los valores para los criterios de restricción pueden ser diferentes para validar los datos buenos y los datos ligeramente contaminados. En esta aplicación se aplicaron restricciones más estrictas al calificar un dato como bueno y menos al calificar un dato ligeramente contaminado. Para considerar un registro como bueno se aceptó un 5% del rango de valores de la curva error, mientras que para calificar un registro como ligeramente contaminado se restringió el criterio de error a un 10%. Para la entropía se utilizó el mismo criterio de restricción en ambos casos, 20% del rango de valores de la curva de entropía. Estos criterios de restricción dependerán de qué estándares de calidad en los datos se requiera.

La Tabla 6.15 presenta los resultados de la clasificación de los registros correspondientes al pozo KU-1001. Se muestran 15 valores del conjunto de datos de este pozo validados. Aparecen la fecha de medición, el diámetro de estrangulador, presión de inyección de BN, gasto de aceite y gasto de gas. Para analizar el comportamiento de la declinación de la producción, se agregó la producción acumulada (N_p) y la presión estática (p_{ws}), que son los datos que se pudieron obtener. Para el caso de la presión estática fue necesario interpolar linealmente los valores que no estaban

disponibles. En las últimas cuatro columnas aparecen la producción acumulada, N_p, el valor de producción diaria, y los valores que se requirieron para analizar el comportamiento de la declinación de este pozo (Q_o /Δp y N_P/ Q_o).En la columna denominada “Calidad”, se presenta el resultado de la validación. Los registros marcados con B fueron los Buenos, los calificados con LC son los ligeramente contaminados, y los señalados con M son los que resultaron malos.

Tabla 6.15. Resultados de la aplicación del método de validación, pozo Ku-1001.

Pozo	Fecha	Estrangulador_inch	PTP_kg_cm2	Piny_kg_cm2	Inyeccion BN_MMpc_d	Qo_bbl_d	Qg_MMpc_d	Calidad	Ps(kg/cm2)	NP bbl	Producción diaria m3	*Y=q/(ps-ptp)*	*X= Np/q*
KU-1001	17-09-03	3.88	16.50	56.70	1.91	6039.00	4.04	LC	136.4	37,440	350.15	49.92	6
KU-1001	19-09-03	3.88	16.20	56.70	2.03	6424.00	4.52	LC	136.4	41,845	350.15	53.06	7
KU-1001	19-09-03	3.88	15.00	56.70	1.53	5943.90	3.94	M	136.4	41,845	350.15	48.65	7
KU-1001	19-09-03	3.88	13.50	56.70	0.95	4986.00	2.63	M	136.1	41,845	350.15	40.34	8
KU-1001	19-09-03	3.88	17.40	56.70	2.51	6328.00	4.97	M	136.1	41,845	350.15	52.96	7
KU-1001	19-06-04	3.88	12.60	53.65	2.03	6596.03	4.52	LC	133.4	1,575,948	964.10	54.01	239
KU-1001	19-12-04	3.88	12.50	55.00	2.00	6429.00	4.46	B	131.7	2,649,954	1036.15	53.46	412
KU-1001	20-08-05	3.88	14.00	68.00	2.74	5440.00	5.04	M	129.3	4,001,832	772.24	46.78	736
KU-1001	15-01-06	3.88	13.40	65.83	2.70	5011.00	4.80	LC	127.8	4,682,899	783.59	43.52	935
KU-1001	27-06-06	3.88	13.10	64.00	2.84	4753.46	4.80	LC	127.3	5,455,602	822.93	41.23	1148
KU-1001	17-07-06	2.25	13.60	65.00	2.89	4188.00	4.56	LC	127.3	5,526,692	519.62	36.61	1320
KU-1001	24-08-06	3.88	15.50	53.00	2.95	5146.00	5.09	M	127.1	5,663,764	626.57	45.83	1101
KU-1001	22-10-06	3.88	16.50	67.00	3.18	4807.00	5.15	M	124.7	6,044,019	904.20	44.06	1257
KU-1001	14-02-07	3.88	14.90	68.00	2.71	4516.00	4.58	M	124.0	6,588,343	776.26	40.96	1459
KU-1001	20-02-07	3.88	15.10	68.70	2.91	4444.00	4.77	B	124.0	6,617,638	776.26	40.38	1489

6.3. Análisis de resultados

La última parte de este trabajo es la validación del modelo, para la cual se debe comprobar que los resultados representen el proceso físico.

La metodología presentada fue aplicada a una base de datos de producción del campo Ku en la formación Cretácico Medio, obteniéndose los resultados siguientes:

- El rango para el parámetro de salida (Q_o, gasto de aceite) durante la validación fue de 1,025 bbld -14,965 bbld.
- Se consideraron tres clases difusas para la clasificación
- La entropía se definió como la razón entre el grado de pertenencia más bajo y el grado de pertenencia más alto. El rango de entropía fue entre 0 y 1.
- La curva de error varió entre 0 y 9500.

Se evaluó el comportamiento de la declinación del pozo KU-1001, tomando los diferentes tipos de registros, buenos, ligeramente contaminados y malos. Para el primer análisis se seleccionaron todos los datos de la Tabla 6.15, sin importar su clasificación. En los análisis realizados en la etapa de preparación de datos se observó que la presión es uno de los parámetros de mayor influencia, por lo que se decidió graficar $Q_o / \Delta p$ contra N_P / Q_o. Además, de acuerdo a la investigación realizada y documentada en el capítulo 3, se encontró que esta es una de las mejores gráficas de diagnóstico para el análisis de los datos de producción (ver **¡Error! No se encuentra el origen de la referencia.**). Debido a las limitaciones que tuvimos para obtener la presión de fondo fluyendo, la caída de presión Δp se obtuvo con la diferencia entre la presión estática y la presión en la tubería de producción. La gráfica del comportamiento de declinación para todos los datos del pozo KU-1001 se muestra en la Figura 6.33; puede notarse que la gráfica de diagnóstico presenta algunos puntos que no son representativos del comportamiento de declinación normal en un pozo.

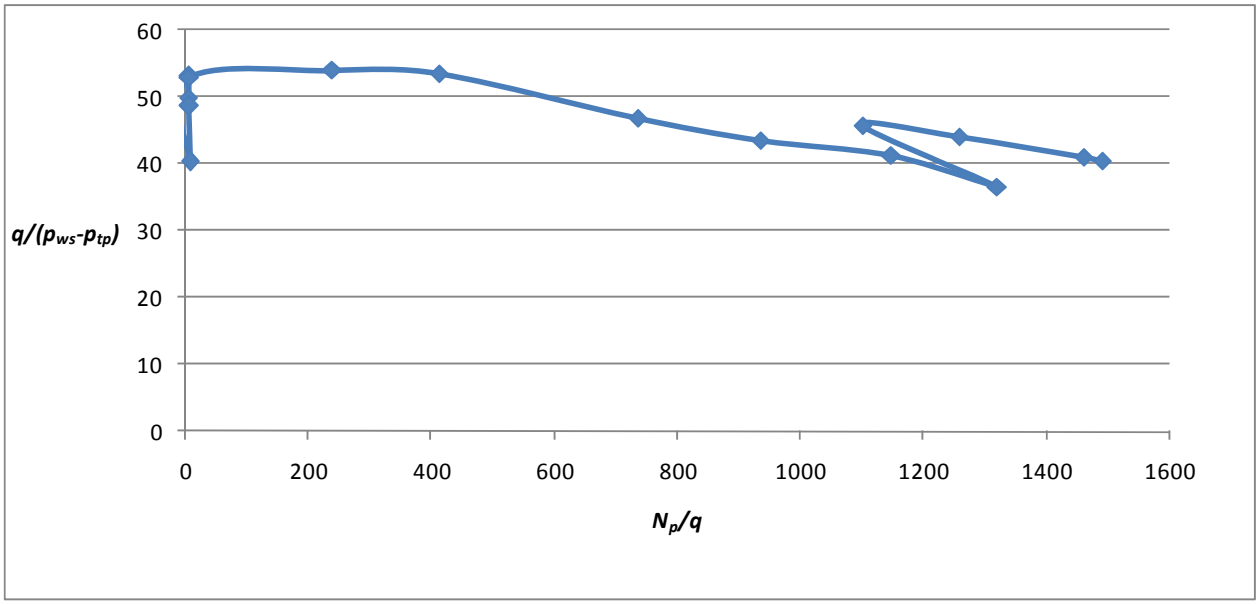


Figura 6.33. Comportamiento de la declinación del pozo KU-1001 incluyendo todos los registros (buenos, ligeramente contaminados y malos)

El paso siguiente en este análisis fue eliminar los registros malos, mostrándose el resultado se muestra en la Figura 6.34, en donde se observa que el comportamiento que se presenta tiene un sentido más físico.

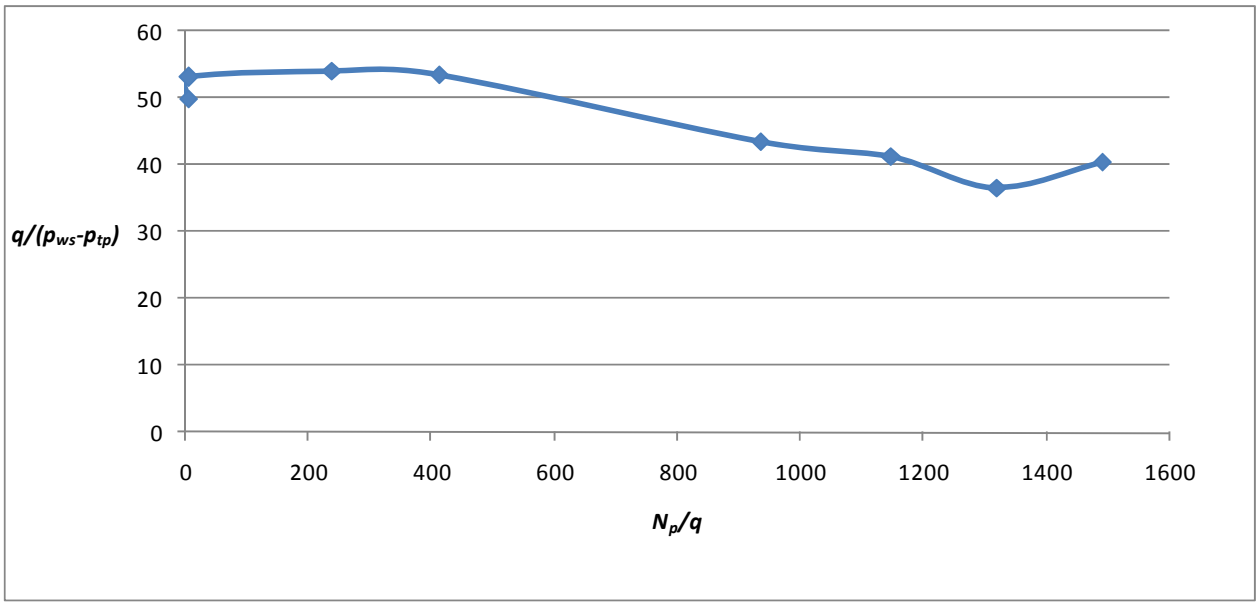


Figura 6.34. Comportamiento de la declinación del pozo KU-1001, empleando datos buenos y ligeramente contaminados.

Finalmente se seleccionaron solo los registros calificados como buenos, los cuales se presentan en la Figura 6.35, en donde para este pozo, siguiendo estrictamente los criterios definidos, hay dos registros calificados como totalmente buenos.

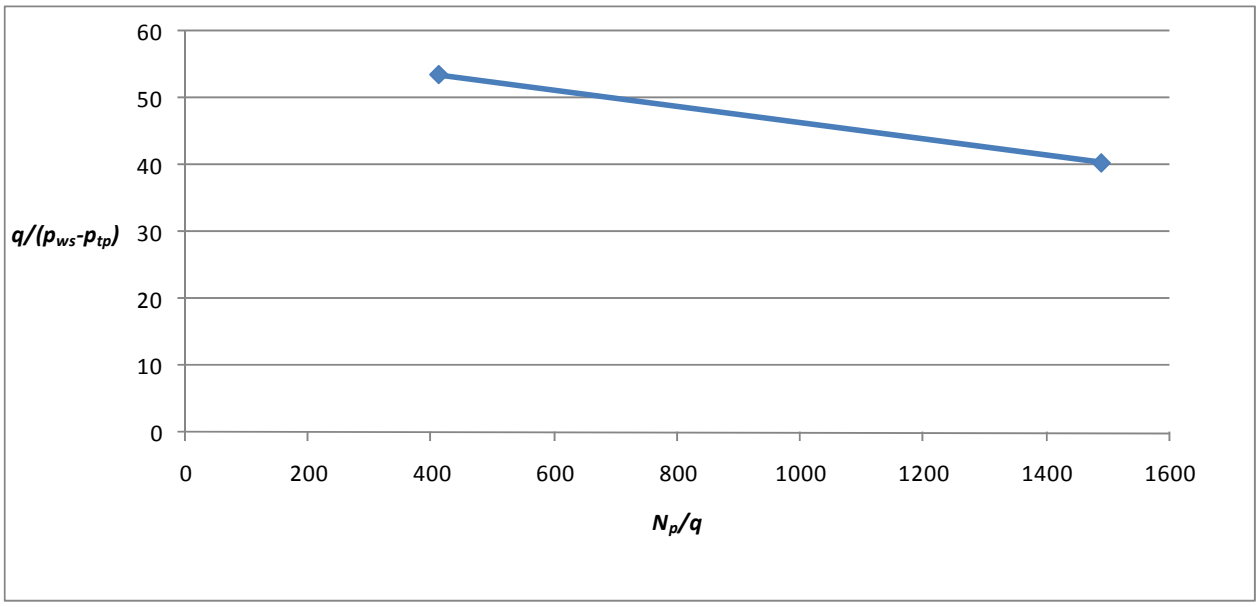


Figura 6.35. Comportamiento de la declinación del pozo KU-1001 considerando registros buenos.

Lo anterior orienta a redefinir los criterios de validación, o a considerar tanto los registros buenos como los ligeramente contaminados, con la calidad suficiente para ser usados en aplicaciones posteriores.

De acuerdo a los resultados presentados en este trabajo, se plantean las siguientes conclusiones y recomendaciones.

CAPÍTULO 7. CONCLUSIONES

Esta metodología demuestra la hipótesis planteada en el Capítulo 5: en un sistema con comportamiento adecuado, la salida deberá ser capaz de contribuir a su propia predicción e identificación.

Se desarrolló un sistema de clasificación Neuro-Cluster para identificar datos contaminados. Se combinaron dos herramientas de inteligencia artificial, redes neuronales y un algoritmo de clasificación difusa, Fuzzy C-Means, lo que generó una solución simple para el problema de clasificación de datos.

El método aplicado es de mucha utilidad, ya que puede aplicarse no solo a datos de producción.

Los resultados indican que hay un cierto grado de subjetividad en el método, ya que al cambiar los criterios de restricción de clasificación, la calidad de los datos es diferente. En el caso analizado se observó que se pueden considerar como válidos tanto los datos buenos como los ligeramente contaminados.

Se pudo analizar la eficacia de las herramientas de minería de datos en la estimación de datos faltantes. El coeficiente de correlación de 0.99 para los datos estimados con respecto a los valores reales, es un indicador importante.

En este trabajo se observó que las limitaciones que introduce la no linealidad de los problemas de ingeniería petrolera pueden superarse con la minería de datos.

El problema de la calidad de los datos puede resolverse en algunas ocasiones por medio del análisis estadístico básico; sin embargo, en ocasiones éste puede ser insuficiente. Problemas que no tienen solución con métodos convencionales pueden tener soluciones muy buenas con este tipo de herramientas.

RECOMENDACIONES

El algoritmo de solución presentado en esta tesis requiere automatizarse, por lo que se recomienda dar continuidad al proyecto en conjunto con personal experto en el desarrollo de herramientas inteligentes, que puedan integrarse en una solución más completa.

En la industria petrolera existen muchas áreas de oportunidad para la aplicación de las herramientas de minería de datos. Por tanto, es recomendable seguir incursionando en este terreno, ya que existen procesos en la ingeniería petrolera altamente no lineales, que requieren una serie de suposiciones al momento de generar sus modelos matemáticos. Esto los hace candidatos viables para modelarse a través de algoritmos de inteligencia artificial.

Algunas aplicaciones importantes de este tipo de herramientas se encuentran en la caracterización de yacimientos: predicción de porosidad, permeabilidad y saturaciones de fluidos, así como en la predicción de gastos de fluidos y presiones de superficie y de fondo, identificación de mejores prácticas y selección de pozos candidatos en procesos de estimulación, así como en procesos de recuperación mejorada.

En resumen, la creatividad, las ideas progresivas y una mente abierta, son ingredientes esenciales en la aplicación de las herramientas de Minería de Datos.

NOMENCLATURA

Términos	Descripción
SOM	= Self Organized Maps, Mapas Auto-Organizados
PDA	= Production Data Analysis, Análisis de Datos de Producción
DCA	= Decline Curve Analysis, Análisis de Curva de Declinación
TCM	= Type Curve Matching, Ajuste por Curva Tipo
RNA	= Red Neuronal Artificial
RA	= Razón de aprendizaje de la red neuronal
AG	= Algoritmo Genético
FCM	= Fuzzy C-Means
BN	= Bombeo neumático
RGA	= Relación gas aceite, m^3/m^3
IDEA	= Intelligent Data Evaluation & Analysis**¡Error! Marcador no definido.**
RGIL	= Relación gas de formación- gas inyección de BN, m^3/m^3
R^2	= Coeficiente de correlación

Variables	Descripción
f_w	= Corte de agua
q_i	= Gasto inicial, volumen/unidad de tiempo
D_i	= Declinación inicial
b	= Exponente de declinación
q	= Gasto, volumen/unidad de tiempo
q_o	= Gasto de aceite, bbl/día
q_g	= Gasto de gas, mmpc/día
q_{inyBN}	= Gasto de gas de inyección de BN mmpc/día
Q_{gform}	= Gasto de gas de formación mmpc/día
Q_{gtotal}	= Gasto de gas total (de formación más de inyección de BN, mmpc/día
p_{wf}	= Presión de fondo fluyendo, kg/cm^2
p_{tp}	= Presión en la tubería de producción, kg/cm^2
p_{baj}	= Presión en el bajante, kg/cm^2
p_{media}	= Presión promedio, kg/cm^2
p_{sep}	= Presión en el separador kg/cm^2
Δp	= Caída de presión
t	= tiempo, unidad de tiempo
T_{sep}	= Temperatura en el separador, °C
T_{baj}	= Temperatura en el bajante °C

P_{sal} = Presión en la salida, kg/ cm^2

p_{iny} = Presión de inyección de BN, kg/ cm^2

p_{inyBN} = Presión de inyección de BN, kg/ cm^2

P_{ws} = Presión estática, kg/ cm^2

N_P = Producción acumulada de aceite mmpc

G_P = Producción acumulada de gas bbl

Variables adimensionales		Descripción
t_{Dd}	= tiempo adimensional	
q_{Dd}	= gasto adimensional	

Subíndices		Descripción
o	= Aceite	
g	= Gas	
w	= Agua	
media	= Promedio	
RNA	= Valor estimado por la red neuronal artificial	
Real	= Valor real	

REFERENCIAS BIBLIOGRÁFICAS

[1] Mohaghegh, S. 2003. Essentials Components of an Integrated Data Mining Tool for the Oil & Gas Industry, With an Application en the DJ Basin. Artículo SPE 84441, presentado en SPE Annual Technical Conference and Exhibition, Denver, Colorado, octubre 5-8. doi: 10.2118/84441-MS.

[1] Mohaghegh, S.D. 2005. Recent Development on Application of Artificial Intelligence in Petroleum Engineering. *J Pet Tech* **57** (4): 86-91. doi: 10.2118/89033-MS.

[1] Popa, A. S., Mohaghegh, S.D., Gaskari, R. and Ameri, S. 2003. Identification of Contaminated Data in Hidraulic Fracturing Databases: Application to the Codell Formation in the DJ Basin. Artículo SPE 83446, presentado en SPE Western Regional/AAPG Pacific Section Joint Meeting, Long Beach, California, mayo 19-24. doi: 10.2118/83446-MS.

[1] Hernández Orallo, J., Ramírez Quintana, M,J, y Ferri Ramírez, C. 2004. *Introducción a la Minería de Datos.* Madrid: Pearson Educación.

[1] Zangl, G. and Oberwinkler, C.P. 2004. Predictive Data Mining Techniques for Production Optimization. Artículo SPE 90372, presentado en SPE Annual Technical Conference and Exhibition, Houston, Texas, septiembre 26-29, 2004. doi: 10.2118/90372-MS.

[1] Zangl, G. and Hannerer, J. 2003. *Data Mining: Application in the Petroleum Industry.* Katy, Texas: Round Oak Publishing.

[1] Osman, E.A., Ayoub, M.A. and Aggour, M.A. 2005. Artificial Neural Network Model for Predicting Bottomhole Flowing Pressure in Vertical Multiphase Flow, Artículo SPE 93632, presentado en SPE Middle East Oil and Gas Show and Conference, Kingdom of Bahrain, marzo 12-15. doi: 10.2118/93632-MS.

[1] Anderson, D.M., Stotts, G.W.J., Mattar, L., Ilk, D. and Blasingame, T.A. 2006. Production Data Analysis: Challenges, Pitfalls, Diagnostics. Artículo SPE 102048, presentado en SPE Annual Technical Conference and Exhibition, San Antonio, Texas, septiembre 24-27. doi: 10.2118/102048-MS.

[1] Gaskari, R., Mohaghegh, S.D. and Jalali, J. 2007. An Integrated Technique for PDA with Application to Mature Fields. *SPE Prod Op* **22** (4): 403-416. doi: 10.2118/100562-PA.

[1] Mohagheg, R., Gaskari, S.D, and Jalali, J. 2005. New Method for Production Data Analysis to Identify New Opportunities in Mature Fields: Methodology and Application. Artículo SPE 98010 presentado en SPE Eastern Regional Meeting, Morgantown, West Virginia, septiembre 14-16. doi: 10.2118/98010-MS.

[1] Arps, J.J.: Analysis of Decline Curves, Trans., AIME(1945),**160**,228-247.

[1] Mattar, L. and McNeil R. 1998. The "Flowing" Gas Material Balance. *J Can Petroleum Technol* **37** (2): 52-55. Disponible en Internet: http://www.fekete.com/resources/papers/flowing_gas_material_bal_paper.pdf [Fecha de acceso 2 de julio de 2010].

[1] Li, K. and Horne, R.N. 2005. An Analytical Model for Production Decline-Curve Analysis in Naturally Fractured Reservoir. *SPE Res Eval & Eng* **8** (3): 197-204. doi: 10.2118/83470-PA.

[1] Blasingame, T.A. and Rushing, J.A. 2005. A Production-Based Method for Direct Estimation of Gas-in-Place and Reserves. Artículo SPE98042, presentado en SPE Eastern Regional Meeting, Morgantown, West Virginia, septiembre 14-16. doi: 10.2118/98042-MS.

[1] Camacho-V., R.G. and Raghavan, R. 1989. Boundary-Dominated Flow in Solution Gas-Drive Reservoirs. *SPE Res Eval & Eng* **4** (4): 503-512. doi: 10.2118/19009-MS.

[1] Palacio, J.C. and Blasingame, T.A. 1993. Decline Curve analysis Using Type Curves –Analysis of Gas Well Production Data. Artículo SPE 25909, presentado en SPE Rocky Mountain Regional/Low Permeability Reservoirs Symposium, Denver, Colorado, abril 12-14.

[1] Doublet, L.E., Pande, P.K., Mccollum, T.J. and Blasingame,T.A. 1994. Decline Curve Analysis Using Type Curves–Analysis of Oil Well Production Data Using Material Balance Time: Application to Field Cases. Artículo SPE 28688, presentado en International Petroleum Conference and Exhibition of Mexico, Veracruz, Ver., octubre 10-13. Disponible en Internet: http://www.pe.tamu.edu/blasingame/data/0_TAB_Public/TAB_Publications/SPE_028688_%28Doublet%29_Material_Balance_Decline_Type_Curve_An.pdf [Fecha de acceso 2 de julio de 2010]

[1] Agarwal, R.G., Gardner, D.C., Kleinsteiber, SW. and Fussel, D.D. 1999. Analyzing Well Production Data Using Combined-Type Curve and Decline-Curve Analysis Concepts. *SPE Res Eval & Eng* **2** (5): 478-486. doi: 10.2118/57916-PA.

[1] Araya, A. and Ozkan, E. 2002. An Account of Decline-Type-Curve Analysis of Vertical, Fractured, and Horizontal Well Production Data. Artículo SPE 77690, presentado en SPE Annual Technical Conference and Exhibition, San Antonio, Texas, septiembre 29 –octubre. doi: 10.2118/77690-MS.

[1] Fuentes-C., G., Camacho-V., R.G. and Vázquez-C., M. 2004. Pressure Transient and Decline Curve Behaviors for Partially Penetrating Wells Completed in Naturally Fractured-Vuggy Reservoirs. Artículo SPE 92116, presentado en SPE International Petroleum Conference in Mexico, Puebla, Pue. noviembre 7-9. doi: 10.2118/92116-MS.

[1] Mattar, L. and Anderson, D. 2003. A Systematic and Comprehensive Methodology for Advanced Analysis of Production Data. Artículo SPE 84472, presentado en SPE Annual Technical Conference and Exhibition, octubre 5-8, Denver, Colorado. doi: 10.2118/84472-MS.

[1] Anderson, D. and Mattar, L. 2004. Practical Diagnostics Using Production Data and Flowing Pressures. Artículo SPE 89939, presentado en SPE Annual Technical Conference and Exhibition, septiembre 26-29, Houston, Texas. doi: 10.2118/89939-MS.

[1] Kabir, CS. and Izgec, B. 2006. Diagnosis of Reservoir Behavior from Measured Pressure/Rate Data. Artículo SPE 100384, presentado en SPE Annual Technical Conference and Exhibition, septiembre 26-29, Houston, Texas. doi: 10.2118/89939-MS.

[1] Bondar, V. 2001. The Analysis of Water-Oil Ratio (WOR) Behavior in Reservoir System, MS dissertation, Texas A&M U., Texas. Disponible en Internet: http://www.pe.tamu.edu/blasingame/data/0_TAB_Grad/TAB_Grad_Thesis_Archive/MS_022A_BONDAR_Valentina_TAMU_Thesis_Vol_1_%28May_2001%29.pdf [Fecha de acceso 27 de julio de 2010].

[1] Nind, T.E.W. 1981. *Principles of Oil Well Production*, second edition. New York: McGraw-Hill.

[1] Fetkovich, M.J. 1980. Decline Curve Analysis Using Type Curves. *J Pet Technol* **32** (6): 1065-1077. doi: 10.2118/4629-PA.

[1] Fetkovich, M.J., Vienot, M.E., Bradley, M.D. and Kiesow, U.G. 1987. Decline Curve Analysis Using Type Curves: Case Histories. *SPE For Eval* **2** (4): 637- 656. Disponible en Internet: http://www.pe.tamu.edu/blasingame/data/z_zCourse_Archive/P689_reference_02C/z_P689_02C_ARP_Tech_Papers_%28Ref%29_%28pdf%29/SPE_13169_Fetkovich_Decline_TC_Case_Histories.pdf [Fecha de acceso 2 de julio de 2010].

[1] Blasingame, T.A. and Lee, W.J. 1986. Variable–Rate Reservoir Limits Testing. Artículo SPE15028, presentado en Permian Basin Oil and Gas Recovery Conference, Midland, Texas, marzo 13-15. doi: 10.2118/15028-MS.

[1] Ilk, D., Valko, P.P. and Blasingame, T.A. 2006. Deconvolution of Variable-Rate Reservoir Performance Data Using B-Splines. *SPE Res Eval & Eng* **9** (5): 582-595. doi: 10.2118/95571-PA.

[1] Ilk, D., Anderson, D.M., Valko, P.P. and Blasingame, T.A. 2006. Analysis of Gas Well Reservoir Performance Data Using B-Spline Deconvolution, Artículo SPE 100573, presentado en SPE Gas Technology Symposium, Calgary, Alberta, Canada, mayo 15-17, 2006. Disponible en Internet: http://www.pe.tamu.edu/blasingame/data/0_TAB_Public/TAB_Publications/SPE_100573_%28Ilk

%29_Analysis_Gas_Well_Res_Perf_Data_Using_B_Spline_Deconvolution.pdf [Fecha de acceso 2 de julio de 2010].

[1] Ilk,D., Mattar, L. and Blasingame, T.A. 2007. Production Data Analysis–Future Practices for Analysis and Interpretation. Artículo PETSOC 2007-174, presentado en Petroleum Society Canadian Institute of Mining, Metallurgy & Petroleum, 8th Canadian International Petroleum Conference (58th Annual Technical Meeting), Calgary, Alberta, Canada, junio 12-14, 2007. Disponible en Internet: http://www.fekete.com/resources/papers/production_data_analysis_future_paper.pdf [Fecha de acceso 2 de julio de 2010].

[1] Smith, J.M., Van Ness, H.C. and Abbott, M.M. 2005. *Introduction to chemical engineering thermodynamics*, seventh edition. Boston, Massachusetts: McGraw-Hill Higher Education.

[1] Acosta Buitrago, M.I. y Zuluaga Muñoz, C. A. 2000. Tutorial sobre Redes Neuronales Aplicadas en Ingeniería Eléctrica y su Implementación en un Sitio Web. Tesis de licenciatura, U. Tecnológica de Pereira, Facultad de Ingeniería Eléctrica, Pereira, Colombia.

[1] Pérez Pueyo, R. 2005. Procesado y Optimización de Espectros Raman mediante Técnicas de Lógica Difusa: Aplicación a la identificación de Materiales Pictóricos. Tesis doctoral, U. Politécnica de Catalunya, Barcelona. Disponible en Internet: http://www.tdx.cat/TDX-0207105-105056/ [Fecha de acceso 2 de julio de 2010].

[1] Mohaghegh, S. 2000. Virtual Intelligence Applications in Petroleum Engineering: Part 3–Fuzzy Logic. *J Pet Technol* **52** (11): 82-87. doi: 10.2118/62415-MS.

[1] Freeman, E. ed. 1983. *The Relevance of Charles Pierce.* La Salle, Illinois: The Hegeler Institute.

[1] Lukasiewicz, J. 1964. *Elements of Mathematical Logic.* New York: MacMillan.

[1] Black, M. 1990. Vagueness: An Exercise in Logical Analysis. *International Journal of General Systems* **17** (2-3): 107-128.

[1] Zadeh, L.A. 1965. Fuzzy Sets. *Information and Control.* **8 (**3): 338-353. Disponible en Internet: http://www-bisc.cs.berkeley.edu/Zadeh-1965.pdf [Fecha de acceso 2 de julio de 2010].

[1] Wei, M., Sung, A. H. and Cather, M. 2004. A Methodology to Discover Contaminated Data in Spatial Databases. Artículo SPE 90267, presentado en SPE Annual Technical Conference and Exhibition, Houston, Texas, septiembre 26-29. doi: 10.2118/90267-MS.

[1] Intelligent Solution Inc, Intelligent Data Evaluation & Analysis. Software de minería de datos con aplicaciones para la industria petrolera. Incluye técnicas de inteligencia artificial tales como redes neuronales, algoritmos genéticos y lógica difusa.

APÉNDICE A
Muestra de datos utilizados en Capítulo 6

Pozo	Fecha	Estrangulador_inch	PTP_kg_cm2	Piny_kg_cm2	Pbajante_kg_cm2	Psep_kg_cm2	Psalida_kg_cm2	TSep_DegC	PinyBN_MMpc_d	Qo_bbl_d	Qg_MMpc_d	QgForm_MMpc_d
KU-1001	2003-09-17	3.875	16.5	56.7	15.4	10	9.1	62	1.91	6039	4.04	2.13
KU-1001	2003-09-19	3.875	16.2	56.7	15.3	9.6	8.6	68	2.03	6424	4.52	2.49
KU-1001	2003-09-19	3.875	15	56.7	14.2	9.4	8.6	66	1.53	5943.9	3.94	2.4
KU-1001	2003-09-19	3.875	13.5	56.7	12.5	9.8	8.6	69	0.95	4986	2.63	1.68
KU-1001	2003-09-19	3.875	17.4	56.7	16.6	10.2	9.1	69	2.51	6328	4.97	2.46
KU-1001	2004-06-19	3.875	12.6	53.651	11.3	10	9.2	83	2.03	6596.03	4.52	2.48
KU-1001	2004-12-19	3.875	12.5	55	11.4	10.68	9.66	80	2	6429	4.46	2.45
KU-1001	2005-08-20	3.875	14	68	13	12.1	11.4	78	2.74	5440	5.04	2.3
KU-1001	2006-01-15	3.875	13.4	65.833	12.7	12.05	11.6	77	2.7	5011	4.8	2.1
KU-1001	2006-06-27	3.875	13.1	64	12	11.4	10.3	78	2.84	4753.46	4.8	1.96
KU-1001	2006-07-17	2.25	13.6	65	12.9	12.4	11.5	69	2.89	4188	4.56	1.67
KU-1001	2006-08-24	3.875	15.5	53	14.8	14.2	13.5	80	2.95	5146	5.09	2.14
KU-1001	2006-10-22	3.875	16.5	67	15.6	14.9	14.1	80	3.18	4807	5.15	1.97
KU-1001	2007-02-14	3.875	14.9	68	13.766667	12.8	11.83	78	2.71	4516	4.58	1.87
KU-1001	2007-02-20	3.875	15.1	68.7	13.9	13.1	12	73	2.91	4444	4.77	1.86
KU-1001	2007-05-17	3.875	13.4	56.6	12.3	11.6	10.6	68	2.73	4953	4.59	1.86
KU-89	2003-02-23	3.25	14.1	66	13.2	12.4	11.7	82.5	1.51	5337	3.87	2.36
KU-89	2003-03-20	3.25	14	66	12.6	12.3	11.5	84	1.71	5372	3.8	2.09
KU-89	2003-06-27	3.25	13.5	61	12.8	12.8	10.6	82	1.03	5024	3.25	2.22