Análisis comparativo entre técnicas de *Machine Learning* para la determinación de unidades de flujo hidráulicas

Jesús Alberto Sierra Angarita

Trabajo de Grado para Optar el Título de Especialista en Ingeniería de Yacimientos

Director

PhD Maika Gambús-Ordaz

Doctora en Ingeniería de Petróleos y Geosistemas

Universidad Industrial de Santander

Facultad de Ingenierías Fisicoquímicas

Escuela de Ingeniería de Petróleos

Especialización en Ingeniería de Yacimientos

Bogotá DC

2023

Dedicatoria

Para mis padres, Jesús y Sayde, quienes a través de su apoyo y esfuerzo incondicional me transmitieron la fuerza y valentía de asumir nuevos retos con proactividad y entusiasmo.

"All models are wrong, but some are useful".

George E. P. Box

Agradecimientos

A la Universidad Industrial de Santander, por abrirme las puertas de la academia pública una vez más y brindarme las herramientas adecuadas para hacer de este programa una experiencia de alta calidad.

A la Dra. Maika Gambús-Ordaz, directora de la investigación, por su apoyo y guía durante el proceso.

A la Agencia Nacional de Petróleo, Gas Natural y Biocombustibles de Brasil, por fomentar el acceso gratuito a datos técnicos públicos de Brasil.

Tabla de Contenido

Pag

Lista de Tablas

Pag

Lista de Figuras

Pag

Resumen

Título: Análisis comparativo entre técnicas de *Machine Learning* para la determinación de unidades de flujo*

Autor: Jesús Alberto Sierra Angarita**

Palabras Clave: Unidades de Flujo Hidráulicas, *Machine Learning*, Indicador de Zona de Flujo, Aprendizaje Supervisado, Aprendizaje No Supervisado.

Descripción: A partir de sitios web brasileros de acceso gratuito, se elaboró una base de datos de registros y análisis de núcleos disponibles, a los que se les realizó un análisis exploratorio de datos encontrando múltiples unidades de flujo bajo la metodología de Amaefule en las formaciones Agua Grande y Sergi. Se aplica el algoritmo no supervisado de mezclas gaussianas para identificar las unidades de flujo a partir de la data de núcleo y determinar sus modelos de permeabilidad al aire a partir de correlaciones con la porosidad efectiva del núcleo y modelos calibrados de porosidad a partir de registros. Finalmente, se aplican algoritmos supervisados y no supervisados en ambos casos de estudio para modelar las unidades de flujo a partir de los registros. Para la formación Agua Grande, se obtienen mejores resultados en algoritmos supervisados, acercándose a 80% de exactitud con el estimador *Gradient Boost Classifier*, mientras que los estimadores no supervisados logran en promedio 60% de exactitud siendo el mejor *K-Means*. Para la formación Sergi, el algoritmo *K-Means* es usado en la identificación de electrofacies, facilitando la interpretación de intervalos gasíferos y arcillosos, demostrando versatilidad sobre los algoritmos de aprendizaje supervisado en ambientes de alta heterogeneidad vertical. Por último, se crearon las plantillas de resultados, incluyendo para Agua Grande el modelo final de las unidades de flujo predichas con sus respectivos modelos de permeabilidad al aire, y para Sergi las electrofacies modeladas en el intervalo de estudio.

* Degree Project

** Faculty of Physicochemical Engineering. School of Petroleum Engineering. Director: Dr. Maika Gambús-Ordaz

Abstract

Title: Comparative analysis between *Machine Learning* techniques for the determination of hydraulic flow units*

Author: Jesús Alberto Sierra Angarita**

Key Words: Hydraulic Flow Units, *Machine* Learning, Flow Zone Indicator, Supervised Learning, Unsupervised Learning.

Description: From Brazilian websites of free access, a database of logs and analysis of available cores was developed, to which an exploratory analysis of data was carried out finding multiple flow units under the methodology of Amaefule in the Agua Grande and Sergi formations. The unsupervised Gaussian mixture algorithm is applied to identify flow units from core data and determine their air permeability models from correlations with effective core porosity and calibrated models of porosity from logs. Finally, supervised and unsupervised algorithms are applied in both case studies to model the flow units from logs. For the Agua Grande formation, better results are obtained in supervised algorithms, approaching 80% accuracy with the *Gradient Boost Classifier estimator*, while unsupervised estimators achieve on average 60% accuracy being the best *K-Means*. For Sergi formation, the *K-Means* algorithm is used in the identification of electrofacies, facilitating the interpretation of gas and clay intervals, demonstrating versatility over supervised learning algorithms in environments of high vertical heterogeneity. Finally, the result templates were created, including for Agua Grande the final model of the predicted flow units with their respective air permeability models, and for Sergi the electrofacies modeled in the study interval.

* Degree Project

** Faculty of Physicochemical Engineering. School of Petroleum Engineering. Director: Dr. Maika Gambús-Ordaz

Introducción

La caracterización de yacimientos es una etapa fundamental para el desarrollo de planes de explotación, ya que permite un correcto entendimiento del potencial petrolífero y productivo del prospecto mediante la evaluación y descripción de sus propiedades geológicas, sedimentológicas y petrofísicas. De esta forma, se pueden interpretar atributos geológicos, texturales, mineralógicos, además de estructuras sedimentarias, contactos y barreras de permeabilidad, junto con propiedades como la porosidad, permeabilidad y presión capilar para la correcta planeación e implementación de estrategias de completamiento en programas de recuperación secundaria y mejorada, al igual que la construcción de modelos representativos de simulación.

En la actualidad, existe una gran variedad de metodologías que estudian la influencia de las variables geológicas mencionadas sobre las variables que controlan el flujo y almacenamiento en la roca, siendo estas principalmente la porosidad efectiva y la permeabilidad absoluta. Una de las metodologías más populares para la caracterización de yacimientos según su condición de flujo y almacenamiento es la propuesta por Amaefule, en donde se definen subdivisiones del yacimiento llamadas unidades de flujo hidráulicas como intervalos dentro de los cuales las propiedades geológicas y petrofísicas que afectan el flujo son consistentes y predeciblemente diferentes de las propiedades de otros volúmenes de roca del yacimiento. Para la implementación de esta metodología, se requieren análisis de núcleos rutinarios donde se pueda evaluar la porosidad efectiva y la permeabilidad al aire y absoluta de la roca, y dado que el costo y tiempo que representa llevar a cabo dicho análisis es alto, en ocasiones no es posible realizar esta caracterización lo suficiente como para comprender correctamente la distribución espacial de las condiciones de flujo y almacenamiento del yacimiento. Ante estas circunstancias, las nuevas tecnologías surgen como

una alternativa de solución para correlacionar los análisis de núcleos con conjuntos de datos mayormente disponible en la evaluación de formaciones como los registros eléctricos, y de esta forma, implementar técnicas de caracterización donde no estén disponibles los núcleos.

En el presente trabajo investigativo, se aplica una metodología basada en técnicas de inteligencia artificial y la metodología de Amaefule para llevar a cabo la caracterización de formaciones geológicas en unidades de flujo de yacimientos siliciclásticos convencionales ubicados en la cuenca del Recóncavo, Brasil. La técnica implementada es llamada *Machine Learning* o aprendizaje automático, la cual es una rama de la inteligencia artificial que se divide en aprendizaje supervisado, no supervisado y reforzado. En los casos de estudio presentes, se implementan técnicas de aprendizaje supervisado y no supervisado únicamente. La principal diferencia entre estos dos tipos de algoritmos consiste en su metodología. En el aprendizaje supervisado se requiere de las etiquetas o resultados de manera previa, siendo estos usados en los modelos de aprendizaje de regresión o clasificación para obtener el resultado deseado, mientras que, en el caso del aprendizaje no supervisado, no se conoce la etiqueta o resultado de manera previa y mediante métodos de *clustering* o agrupamiento se busca identificar patrones de características similares para categorizarlos y obtener el resultado deseado.

Teniendo esto en cuenta, en el capítulo 1 de la investigación se enlistan los objetivos generales y específicos que demarcarán el flujo de trabajo. En el capítulo 2 se desarrolla el marco de referencia, en donde se incluyen los antecedentes investigativos de la tecnología usada y el marco teórico conceptual donde se contextualiza el área de estudio y se definen los principales conceptos de la investigación. Luego, en el capítulo 3 se realiza el análisis exploratorio y estadístico de datos requerido al inicio de todo proceso de aprendizaje automático, partiendo de la recopilación de datos de acceso gratuito en la web, seguido de la identificación de las formaciones

presentes, tipo y recuento de datos y análisis estadístico de estos, para posteriormente seleccionar el conjunto de datos a entrenar en el modelo y plantear los casos de estudio respectivos según las formaciones a analizar. En el capítulo 4 se realiza la identificación de las unidades de flujo hidráulicas bajo la metodología de Amaefule y se establecen los modelos de porosidad total, porosidad efectiva y permeabilidad al aire para cada unidad según la data disponible reportada. Posteriormente, en el capítulo 5 se aplican técnicas de aprendizaje supervisado y no supervisado con el conjunto de datos establecido para los casos de estudio planteados, reportando sus métricas de rendimiento. Finalmente, en el capítulo 6 se expone el análisis de resultados, en donde se escalan los modelos obtenidos del capítulo anterior sobre la totalidad del intervalo de interés de cada caso de estudio y se comparan sus resultados, creando por último un *template* con los registros de pozo, los análisis de núcleos, modelos de porosidad efectiva y permeabilidad al aire, y las unidades predichas en el intervalo de estudio con el mejor modelo establecido.

1. Objetivos

1.1 Objetivo General

Realizar un análisis comparativo entre técnicas de aprendizaje automático para la determinación de unidades de flujo en formaciones geológicas de la cuenca Recóncavo, Brasil.

1.2 Objetivos Específicos

- Analizar exploratoria y estadísticamente el conjunto de datos adquiridos a través de los registros de pozos y núcleos disponibles de las formaciones geológicas en estudio.
- Establecer los modelos de porosidad efectiva y permeabilidad mediante el uso de registros de pozos calibrados con datos de núcleos en la evaluación de las propiedades petrofísicas.
- Determinar las unidades de flujo presentes mediante algoritmos de aprendizaje supervisado (regresión y clasificación) y aprendizaje no supervisado (clustering o agrupamiento).
- Comparar los resultados de las unidades de flujo identificadas a partir de las diferentes técnicas de inteligencia artificial aplicadas.

2. Marco de Referencia

2.1 Antecedentes Investigativos

Dada la importancia y relevancia que tiene la caracterización de yacimientos en la descripción de las propiedades de almacenamiento y flujo de hidrocarburos, desde hace más de 40 años se han estado presentando propuestas en el área con el fin de solucionar inconvenientes relacionados a la falta de información o competencias al momento de realizar análisis petrofísicos convencionales. En la década de 1980 se introdujo el concepto de agrupamiento y clustering de atributos similares en la clasificación de facies, cuando en aquel entonces se solían usar métodos con enfoque estadístico multivariado para dichos análisis. Por ejemplo, Delfiner et al. (1987) y Busch et al. (1987) aplicaron una función de análisis discriminante para estimar facies. Gill et al. (1993) usó un clustering multivariado y correlación de zonas entre pozos para determinar facies. Debido al continuo avance en el desarrollo del *Machine Learning*, los esfuerzos investigativos eran cada vez mayores en su aplicación para la clasificación de facies. Baldwin et al. (1990) aplicó redes neuronales para identificar minerales en registros de pozos. Rogers et al (1992) y Kapur et al. (1998) usaron redes neuronales para predecir facies de núcleos y registros de pozos.

En los últimos años, la aplicación de técnicas de inteligencia artificial se ha visto notablemente en aumento en toda la cadena de valor de la industria del petróleo. Por ejemplo, Kadkhodaie-Ilkhchi et al. (2013), estudiaron la aplicación del clustering en la identificación de electrofacies y su relación con litofacies y unidades de flujo previamente identificadas. Para esta integración, usaron los registros eléctricos junto con datos de permeabilidad y porosidad en dos enfoques: sedimentario y petrofísico. En el enfoque sedimentario, las electrofacies fueron

identificadas según su textura sedimentaria y propiedades litológicas, mientras que, en el enfoque petrofísico, las electrofacies fueron identificadas de acuerdo con las propiedades petrofísicas. De forma adicional, para encontrar una buena conexión entre las electrofacies y zonas productoras, se incorporó el concepto de unidades de flujo hidráulicas. Para la determinación de electrofacies bajo el enfoque sedimentario, se realizó el clustering usando los registros y en particular el GR como discriminante en apoyo de estudios petrográficos y descripciones de los núcleos, mientras que, en el enfoque petrofísico, se les dio más peso a los registros de porosidad y densidad. De esta forma, identificaron 4 electrofacies petrofísicas (EF) y 5 unidades de flujo (HFU). Una de las notas relevantes del estudio es la presencia de más de una HFU por cada EF y la variedad de las propiedades sedimentarias dentro de una misma EF. Por ejemplo, la EF1 está compuesta principalmente por areniscas de grano medio a grueso y muy grueso relacionadas a las HFU E y D, sin embargo, en esta EF se tiene una pequeña presencia de areniscas finas relacionadas a las HFU C y D. Una primera conclusión de este trabajo investigativo fue la flexibilidad que demostró la técnica en la caracterización del yacimiento, ya que permite integrar y discriminar categorías como HFU y EF como distribuciones de probabilidad, evidenciando que las unidades de flujo hidráulicas y las electrofacies identificadas no coinciden necesariamente con el tipo de roca del yacimiento o sus facies. La tabla 1 resume los resultados de la investigación.

Khalid et al. (2019), desarrollaron una metodología basada en un análisis de regresión múltiple para predecir la permeabilidad en pozos no corazonados a partir de la identificación de unidades de flujo y su relación con los registros eléctricos. Así mismo, Shi et al. (2019) describieron un flujo de trabajo novedoso que predice de forma continua la permeabilidad a partir de registros convencionales, basado en la clasificación de electrofacies y análisis de núcleos recolectados en múltiples campos de petróleo. Desarrollaron una técnica llamada *Multi-Resolution*

Graph based Clustering (MRGC) usada para clasificar electrofacies de las curvas de registros en las secciones corazonadas. Luego, usaron el algoritmo KNN para entrenar los resultados de la clasificación de electrofacies en secciones no corazonadas. Finalmente, el modelo de permeabilidad basado en la condición de electrofacie es establecido, calculando además el índice de productividad de los pozos seleccionados.

Tabla 1

Descripción cuantitativa de los resultados por Kadkhodaie-Ilkhchi et al. (2013)

Tight sand rock typing in the Whicher Range Field

Rock type	DT	RHOB	GR	Φ (FR)	K (MD)	EF	HFU	Depositional facies
Very tight sandstone								
MIN	58.4	1.8	40.1	0.01	0.01	1	D and E	Medium/coarse to very coarse sand, tightly cemented by calcite, silica and clay cements
AVG	62.4	2.3	69.5	0.03	0.53			
MAX	64.3	2.6	91.2	0.06	3.93			
STDEV	1.6	0.17	15.5	0.01	0.66			
Tight sandstone								
MIN	64.6	1.6	40.3	0.03	0.007	2	D and E	Medium to coarse sand, cemented by silica, calcite and clay, poor to fair intergranular porosity
AVG	67.9	2.3	79.7	0.06	0.87			
MAX	70.9	2.7	131.4	0.1	4.62			
STDEV	1.7	0.21	23.8	0.02	0.96			
Sub-tight sandstone								
MIN	71.1	1.6	37.4	0.06	0.19	3 and 4	C and D	Medium to coarse sand, affected by silica, calcite and clay cements, poor to fair intergranular porosity
AVG	75.8	2.1	62	0.13	1.65			
MAX	87.2	2.6	124.9	0.22	7.87			
STDEV	4.2	0.18	22.4	0.03	1.66			
Fine grained and silty sandstone								
MIN	64.6	1.5	40.6	0.04	0.001	1–4	A and B	Very fine/fine and argillaceous sand
AVG	74.5	2.2	87.9	0.12	0.09			
MAX	97.4	2.6	131.3	0.27	0.8			
STDEV	6.9	0.29	25.87	0.05	0.16			

Nota. Tomado de Kadkhodaie-Ilkhchi et al. 2013

Según Abbas et al. (2019) en los últimos años, se ha intensificado la investigación y aplicación de técnicas de aprendizaje no supervisado en el reconocimiento de patrones de identificación de litofacies y electrofacies a partir de datos de núcleos y registros de pozo, con el objetivo de predecir las facies presentes en intervalos donde no se han realizado operaciones de corazonamiento y así poder realizar una mejor estimación de las propiedades de capacidad de flujo de un intervalo. En su caso particular, llevaron a cabo la comparación de técnicas convencionales de clasificación de facies como el indicador de zona de flujo y unidades de flujo hidráulico, con

modelos de aprendizaje no supervisados basados en el análisis de Clusters, como el modelo K-means e Hierarchical Clustering. Los resultados obtenidos mostraron al modelo K-means como el más preciso en comparación con los datos reales de pozo y cumplió con gran acierto en sus predicciones de clasificación de facies.

Fadokun et al. (2020) plantearon dos enfoques de aprendizaje automático para la predicción de facies usando el lenguaje Python. En el enfoque no supervisado, usaron la técnica de clustering la cual involucra el agrupamiento de datos basado en similitudes y distancias, y en el supervisado la técnica *support vector machine* en donde se establece una data objetivo (data de núcleo). Para garantizar el éxito de sus modelos, emplearon técnicas de limpieza, preprocesamiento y visualización de datos. Determinaron que el primer paso y más importante al crear un modelo predictivo es la preparación de los datos, lo cual involucra la limpieza y el preprocesamiento. En el estudio realizado por Fadokun et al. (2020), usaron el enfoque de estandarización para esta tarea, el cual es un proceso de re-escalamiento de uno o más atributos para que tengan una media de cero (0) y una desviación estándar de uno (1). Para el enfoque de aprendizaje no supervisado, usaron previo al modelo la técnica *Principal Component Analysis* para reducir la dimensionalidad de los datos con el fin de convertir un *set* de variables posiblemente relacionadas en un set de variables linealmente no correlacionadas, es decir, componentes principales. El algoritmo usado para el aprendizaje supervisado *support vector machine* (SVM) es un modelo que analiza la clasificación de la data involucrada y problemas de regresión. El objetivo del SVM es encontrar un hiperplano con el margen más largo de cada punto de observación graficado para discriminar y clasificar estos puntos como clases distintas.

Hong et al. (2020) desarrollaron un modelo de identificación de facies no supervisado basado en redes neuronales. Para el algoritmo, se aplicaron *clusters* no supervisados según las

similitudes en las respuestas de las rocas a los distintos registros geofísicos. De las tantas técnicas existentes para el desarrollo de algoritmos de aprendizaje no supervisados, Hong et al. aplicaron la denominada “Información que maximiza el entrenamiento autoaumentado” y realizaron la comparación de su modelo con las técnicas populares de K-means clustering, Spectral Clustering, entre otras.

Recientemente, Robail et al. (2023) usaron un enfoque de *machine learning* para soportar el estudio integrado multidisciplinar de la construcción de un modelo de campo a escala de yacimiento, el cual había adquirido datos de núcleos adicionales para varios pozos recientemente perforados, siendo descritos por sedimentólogos definiendo las facies depositacionales del yacimiento y las litofacies. Esta descripción fue usada por un algoritmo de *machine learning* para entrenar los registros convencionales *triple combo* con el fin de reconocer las facies del yacimiento. Luego, estas facies geológicas fueron propagadas usando registros en más de 80 pozos sin corazonar, realizando predicción de las facies presentes en un contexto geológico. Robail et al. (2023) obtuvieron una replicación excelente en los pozos corazonados, así como resultados robustos en pozos no corazonados del campo. Pudieron verificar la robustez del modelo también con registros de producción e inyección (PLT/ILT), tomografía computarizada y descripción de núcleos. Las facies geológicas predichas en pozos corazonados y no corazonados fueron usadas junto con las tendencias de inversión sísmica para condicionar la distribución 3D de las facies en el modelo del yacimiento. El uso de *machine learning* para la predicción de facies les ayudó también a validar el concepto geológico subyacente de un intervalo de yacimiento antiguo de buena calidad en ciertas áreas del campo, las cuales no fueron correctamente muestreadas con los núcleos existentes. Por último, establecieron recomendaciones futuras para usar modelos de yacimiento

basados en *machine learning* en la identificación de nuevas ubicaciones *infill* donde las mejores facies productoras estén predominantemente presentes.

2.2 Marco Teórico Conceptual

2.2.1 Sistema Petrolífero Cuenca Recóncavo

La Cuenca del Recóncavo, objetivo de la presente investigación, está ubicada en la porción este del Estado de Bahía, región nordeste de Brasil. Cubre un área de aproximadamente 11000 km^2 y presenta una orientación general que sigue la tendencia NE-SW. Se limita al norte y al noroeste con la Cuenca del Tucano, por el alto de Aporá; al sur con la Cuenca de Camamu, a través del sistema de fallas de Barra; al este, por el sistema de fallas de Salvador; y al oeste por la falla de Maragogipe.

El origen de la Cuenca del Recóncavo está ligado al proceso de estiramiento de la corteza que, durante el cretácico inferior, resultó en la fragmentación del continente Gondwana y la apertura del Océano Atlántico sobre el cratón de San Francisco. La cuenca compone el conjunto de depósitos cretácicos a lo largo de la costa este de Brasil. Constituye el segmento de una grieta intracontinental abortado y su arquitectura básica refleja un semi-graben con orientación NE-SW, donde la falla del borde este presenta relieves eventualmente mayores a 6 km. El basamento precámbrico de la cuenca está formado por rocas arcaicas del Paleoproterozoico, perteneciente al bloque Serrinha y los cinturones de Itabuna Salvador-Curacá, y por rocas metasedimentarias de

edad Neoproterozoico perteneciente al grupo Estancia. Se estima que la sección sedimentaria conservada en la Cuenca del Recóncavo tiene un espesor del orden de 6900 metros (Prates, 2017).

Figura 1

Cuenca del Recóncavo, Brasil

Nota. La figura muestra la ubicación del límite y marco estructural de la Cuenca del Recóncavo. Tomado de Prates, 2017.

Debido al control que la actividad tectónica ejerció sobre la sedimentación, la depositación de la cuenca se desarrolló en tres fases principales, correspondientes a tres supersecuencias representadas en el enfoque estratigráfico de la figura 2.

La fase previa al *rift* representa la primera supersecuencia estratigráfica, depósitos relacionados con la etapa inicial de la flexión de la corteza terrestre y se extienden desde el Neo-Jurásico al cretácico inferior. Esta secuencia está compuesta por shales rojos y areniscas de la

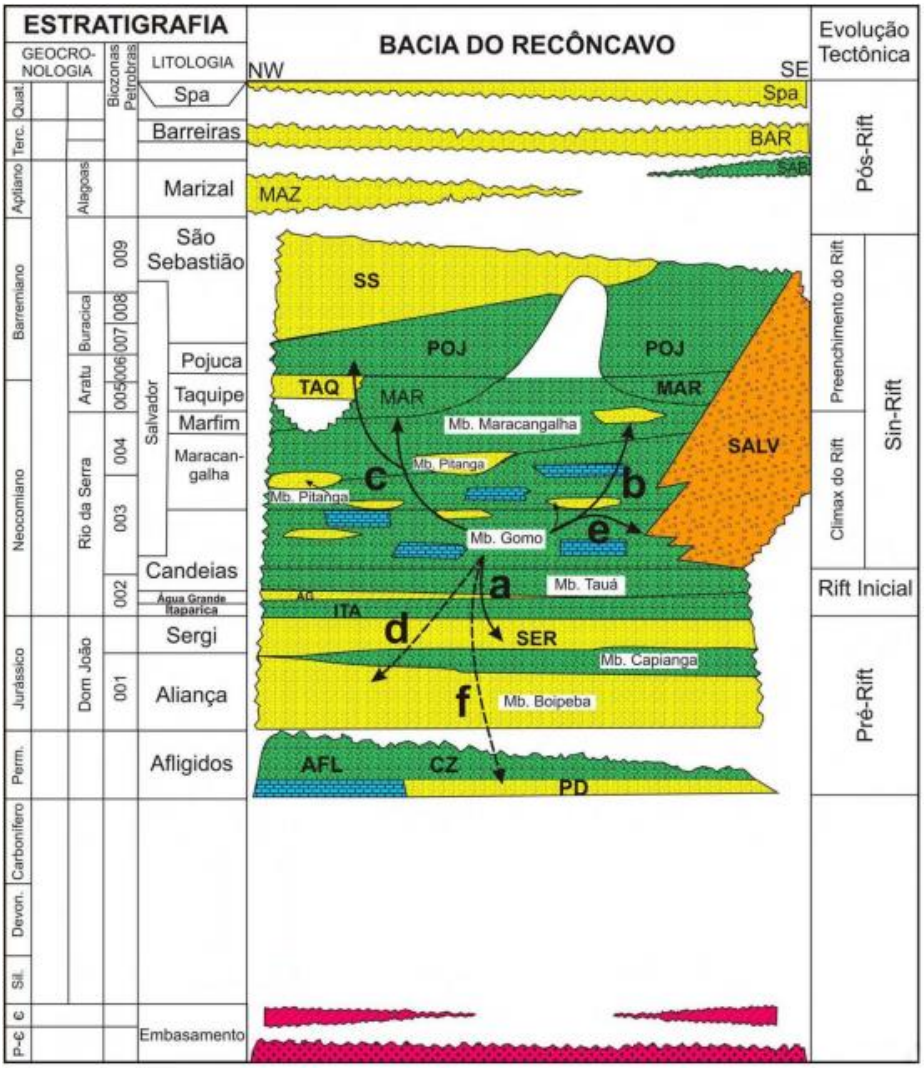
Formación Alianca superpuesta por ciclos fluvio-eólicos representados las formaciones Sergi y Agua Grande. Estos ciclos son separados por transgresiones lacustres representadas por sedimentos del Miembro Capianga de la Formación Alianca y por sedimentos de la Formación Itaparica. La supersecuencia que corresponde a la fase de *rift* comenzó con el aumento de la tasa de hundimiento y un cambio climático repentino. Cuando nuevamente se implantó un sistema lacustre, anóxico e inicialmente somero, se depositaron los sedimentos del Miembro Tauá de la Formación Candeias, cuyo límite marca el comienzo de una intensa tafrogenia. La depositación de sedimentos arcillosos intercalados con carbonatos (Miembro Gomo de la Formación Candeias) ocurrió debido a la formación de lagos profundos del proceso de tafrogénesis. Entre estos y la Formación Sergi, los *shales* lacustres de la formación Itaparica, dividida por diques de arena, y las areniscas fluviales a eólicas de Agua Grande proveen una transición entre las secuencias *prerift* y *synrift*. Con la expansión y profundización de la cuenca, tuvo lugar la sedimentación de la Formación Maracangalha, similarmente compuestas de shales lacustres y areniscas turbidíticas, pero en aguas menos profundas.

Durante las etapas tardías de la evolución del *rift*, estos *shales* fueron fuertemente deformados por el peso de los sedimentos superpuestos, formando diapiros de shale que penetraron la secuencia del *rift*. Las formaciones Candeias y Maracangalha juntas forman el grupo Santo Amaro (tabla 2).

Después de la depositación de los *shales* y turbiditas del grupo Santo Amaro, el flujo de sedimentos a la cuenca incrementó mientras que la subsidencia disminuyó, lo que inició el llenado de la cuenca. Dos grandes cuerpos sedimentarios se formaron: los conglomerados del Salvador, compuesto por sedimentos abanico-deltaicos derivados del horst Salvador-Jacuípe; y los sedimentos del grupo Ilhas, en donde dominan areniscas delticas (formaciones Marfim y Pojuca).

Figura 2

Carta estratigráfica de la Cuenca del Recóncavo



Nota. Carta estratigráfica de la Cuenca del Recóncavo. Las flechas curvas identificadas como (a), (b), (c), (d), (e) y (f) representan rutas probables para la migración de hidrocarburos desde la roca generadora principal. Tomado de CPGG-UFBA, 2008.

Tabla 2

Tabla estratigráfica de la Cuenca Recóncavo

Period	Epoch/Age	Group	Formation	Member	Environment	Lithology (main)
Cenozoic	Pliocene	Barreiras			Alluvial fan	Sandstone
L. Cretaceous (POST-RIFT)	Aptian		Marizal		Alluvial fan	Conglomerate
(SYN-RIFT)	Barremian	Massacará	Poço Verde			Shale
			São Sebastião		Fluvial	Sandstone
	Hauterivian	Ilhas	Pojuca	Santiago	Deltaic	Sandstone/ Shale
			Taquipe		Canyon fill	Sandstone/ Shale
	Valanginian		Marfim	Catu	Deltaic	Sandstone
			Salvador	Sesmaria	Fan-delta	Conglomerate
		Santo Amaro	Maracangalha	Pitanga/ Caruaçu	Lacustrine/ Turbidites – Sand debris	Shale/ Sandstone
	Berriasian		Candeias	Gomo	Lacustrine/ Turbidites	Shale
				Tauá	Lacustrine	Shale
(PRE-RIFT)			Água Grande		Fluvial/aeolian	Sandstone
U. Jurassic			Itaparica		Lacustrine	Shale
	Tithonian	Brotas	Sergi		Fluvial/aeolian	Sandstone
			Aliança	Capianga	Playa lake	Shale
				Boipeba	Fluvial/aeolian	Sandstone
Permian			Afligidos/ Santa Brígida	Cazumba/ Ingá	Playa lake	Shale
	Kungurian			Pedrão/ Caldeirão	Platform/tidal	Sandstone

Nota. Tomado de Magnavita et al., 2012

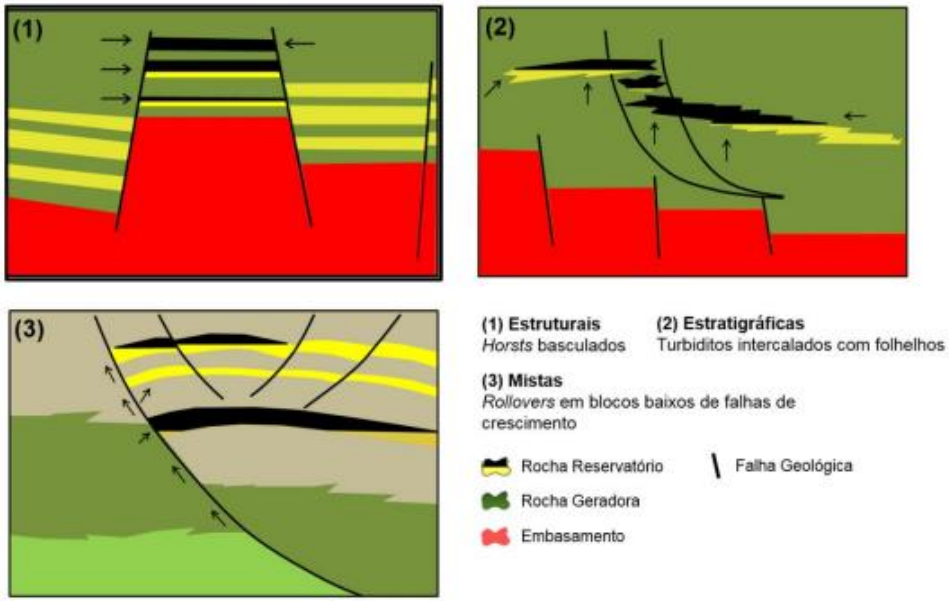
Según Prates (2017), existen tres modelos primarios de migración y acumulación en la cuenca Recóncavo: (1) trampas estructurales formados por *horsts* inclinados o no, donde los depósitos *prerift* se alimentan lateralmente desde la generación de lutitas ubicada en los bajos de fallas tensionales, (2) trampas estratigráficas, principalmente yacimientos turbudíticos de las formaciones Candeias y Marfim, conectadas directamente a las lutitas generadoras con distancias

de migración cortas, y (3) *rollovers* formados a lo largo de fallas de crecimiento de la secuencia *synrift*, a nivel de los embalses deltaicos de las formaciones Pojuca y Marfim con migración vertical a través de fallas regionales (figura 3).

Esta descripción concuerda con la expuesta por Mello et al. (1994), donde ejemplifica el tipo de trampas petroleras con dos de los campos petroleros más grandes de la cuenca que juntos contienen 1152 millones de barriles: Agua Grande que representa trampas *prerift* y *synrift* incluyendo secuencias depositacionales en la Formación Candeias inferior, y Miranga que representa trampas synrift incluyendo secuencias depositacionales en la Formación Candeias superior y el grupo Ilhas (figura 4).

Figura 3

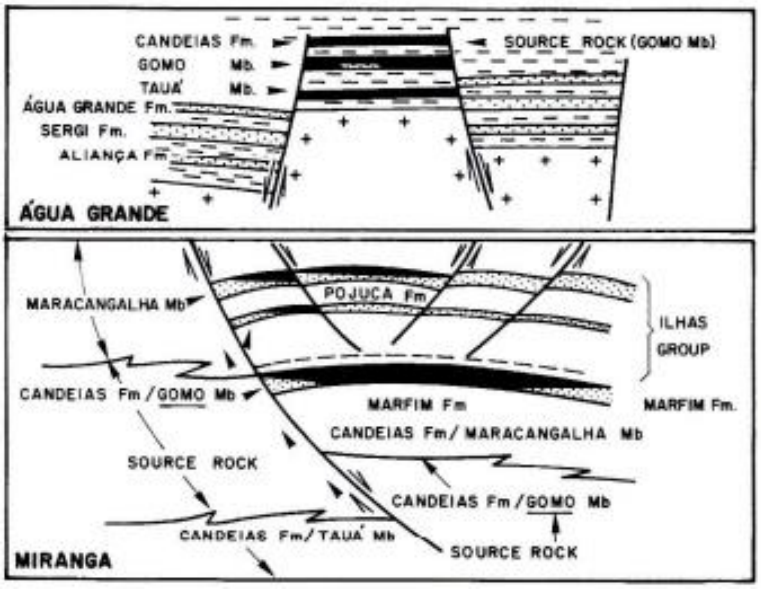
Modelos de migración y acumulación de la Cuenca Recóncavo



Nota. Tomado de Prates, 2017.

Figura 4

Secciones transversales de los campos (a) Agua Grande y (b) Miranga



Nota. Tomado de Mello et al., 1994

La Cuenca Recóncavo es la cuenca de mayor antigüedad en producción en Brasil y se encuentra en fase madura de exploración. La producción comercial inició en 1941 y producto de inversiones masivas en las décadas siguientes se han logrado descubrir 85 campos de petróleo y gas.

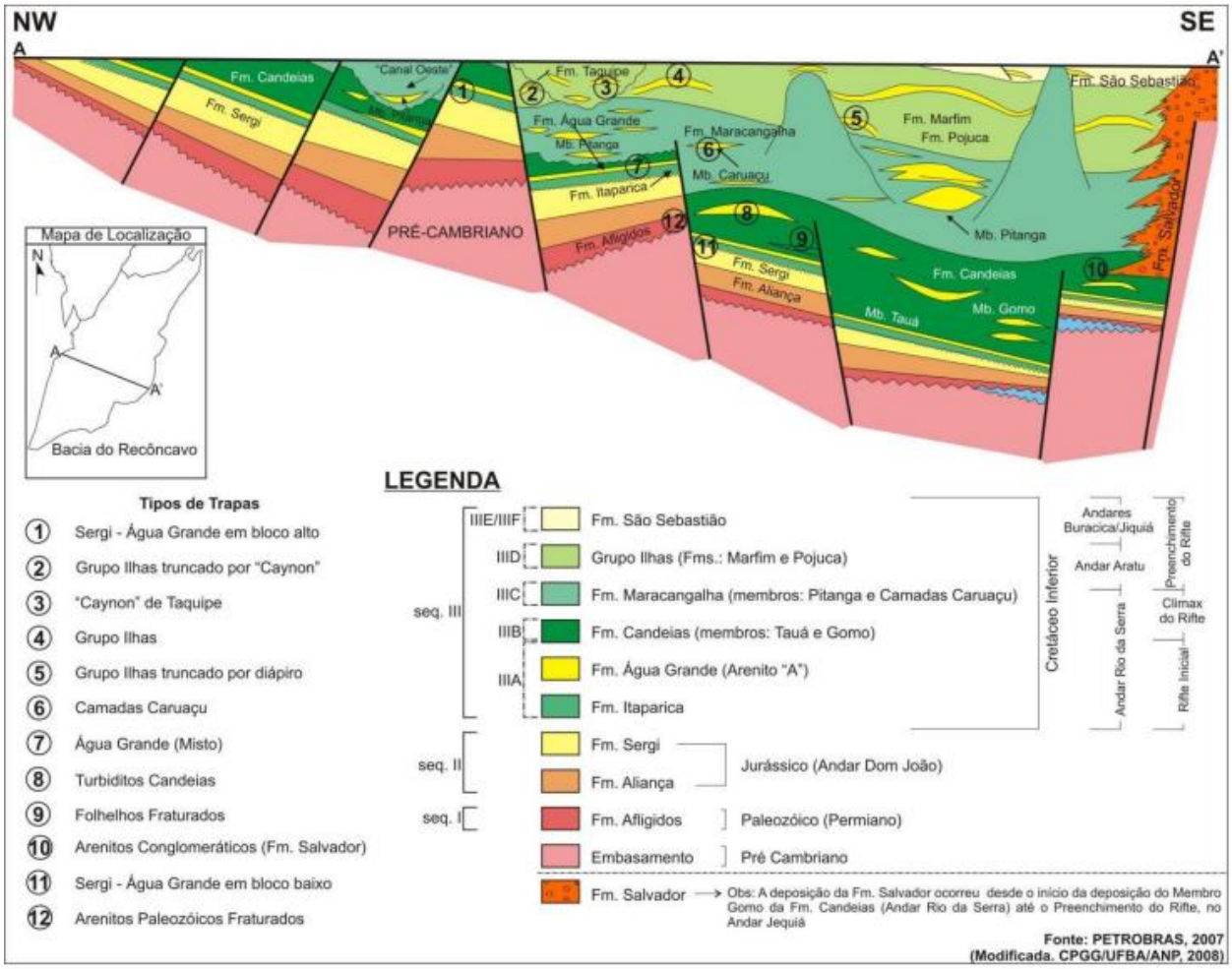
La acumulación de hidrocarburos en la cuenca puede agruparse en cuatro sistemas básicos, en orden de importancia: (1) el sistema *prerift* Candeias-Sergi/Agua Grande, (2) el sistema *synrift* Candeias-Ilhas, (3) el sistema *synrift* Candeias/Candeias, y (4) el sistema *synrift* Candeias/Caracu (miembro de la formación Maracangalha).

De esta forma, la principal roca generadora en la cuenca se establece como *shale* lacustre en los miembros Tauá y Gomo de la Formación Candeias, con valores de TOC que pueden llegar hasta 10%, kerógeno tipo 1 y potencial de generación de hidrocarburos de 5 kg HC/ton roca. La ventana de petróleo coincide con las principales estructuras bajas, y análisis geoquímicos muestran

que más del 80% del petróleo fue expulsado (Magnavita et al., 2012). A su vez, la formación Pojuca tiene un potencial moderado generador, pero ocurre por encima de la ventana de generación en la mayor parte de la cuenca, con un límite superior de 2400 metros de profundidad.

Figura 5

Sección geológica esquemática de la Cuenca del Recóncavo



Nota. En la imagen se puede apreciar la presencia de los diferentes tipos de trampas mencionadas anteriormente, así como observar los elementos cronoestratigráficos y la división de secuencias depositacionales. Tomado de CPGG-UFBA, 2008.

Los principales yacimientos de la cuenca están compuestos por areniscas eólico-fluviales de las formaciones Sergi, Itaparica y Agua Grande, turbiditas de las formaciones Candeias y areniscas de Maracangalha y fluvio-deltaicas de las formaciones Marfim y Pojuca. Las principales roca sello son sedimentos finos (arcillosos): lutitas de miembros Tauá y Gomo de la Formación Candeias, lutitas de la Formación Maracangalha, lutitas prodeltaicas de las formaciones Marfim y Pojuca y las lutitas de la Formación Taquipe. Para el principal sistema petrolífero de la cuenca, las rocas sello son las lutitas de las formaciones Itaparica y Candeias (Prates, 2017).

2.2.2 Machine Learning

Según Narayan et al. (2020), el *Machine Learning* es el campo de estudio en el cual los computadores son algorítmicamente programados para aprender y adaptarse desde la experiencia para mejorar en una tarea de evaluación de procesos usando métricas. La etapa de emplea grandes volúmenes de información para permitir al computador el reconocimiento de patrones ocultos. Este campo se puede clasificar en distintas categorías:

- Aprendizaje supervisado: En este aprendizaje, el resultado deseado es conocido y el algoritmo de Machine Learning provee un esquema general entre los datos de entrada y las variables de salida. Las dos subcategorías principales del aprendizaje supervisado son la regresión y la clasificación de problemas, los cuales están definidos según el tipo de variable de salida. En el aprendizaje supervisado, el proceso de modelaje de entrenamiento continua con la evaluación del error y haciendo mejoras hasta que el nivel deseado de precisión es alcanzado.

- Aprendizaje no supervisado: En el aprendizaje no supervisado, no hay una variable explícita de salida, y las relaciones son generadas basadas en la información suministrada al algoritmo. Algunos de los algoritmos que pertenecen a esta categoría pueden revelar estructuras escondidas y relaciones entre datos de entrada. Algunos ejemplos incluyen clustering, algoritmos de reducción dimensional, y regla de aprendizaje asociativo.

Según los autores, (Nayara et al. 2020), existen una gran variedad de lenguajes de programación que son usados en el desarrollo de soluciones y aplicaciones de *Machine Learning*, por ejemplo, Python, R, Java, C/C++, Julia, Scala, Go y Lua, son de los más populares. En el caso particular de Python, representa en lenguaje con la mayor comunidad de practicantes de Machine Learning desde 2019, y a pesar de que cada lenguaje tiene ventajas y desventajas, Python se posiciona en el mercado como un ecosistema rico en herramientas para la aplicación del *Machine Learning*. Algunas de las librerías más populares de Python para aplicaciones de *Machine Learning* son por ejemplo Scikit-Learn, la cual es una construcción de otras librerías como NumPy, Scipy, y Matplotlib, TensorFlow, Keras, Theano, PyTorch, OpenCV.

Para construir un modelo de *Machine Learning*, se necesita primero el preprocesamiento de la información, lo cual incluye los pasos de normalización, estandarización y escalamiento, según el algoritmo seleccionado. Luego de esto, se realiza la división de datos, que incluye el entrenamiento, validación, y testeo.

Para el aprendizaje no supervisado, como se mencionó anteriormente, no hay asociada una variable de salida, sino que el algoritmo trabaja para identificar patrones y relaciones escondidas en la información suministrada. Uno de los métodos más populares para la elaboración de algoritmos en esta categoría es el clustering, el cual consiste en encontrar clusters o grupos de

características similares en la información de muestra. Los algoritmos de clustering trabajan muy bien sobre la muestra, donde grupos distintivos de la muestra están presentes. Una vez que el algoritmo de clustering es entrenado, cualquier nueva observación es predicha a pertenecer a alguno de los clusters identificados. En particular, el algoritmo de clustering k-means es usado para segmentar un grupo de información con n observaciones en k clusters distintivos centrados en un centroide. En la presente investigación se propone el desarrollo de varias técnicas de *Machine Learning* integradas en aprendizaje supervisado y no supervisado para la generación de modelos de predicción de unidades de flujo y ser usados en pozos sin núcleo, de manera que con modelos validados se puedan obtener estas unidades con menos recursos.

El tipo de algoritmos supervisado por lo general suelen dar los mejores resultados de precisión con respecto a los valores originales. De manera alterna, se desarrollará un algoritmo de aprendizaje no supervisado, en donde a diferencia del anterior, no se presentarán las unidades de flujo previamente identificadas, y se espera que únicamente del comportamiento de los registros se puedan identificar. Este tipo de algoritmos suelen tener menos precisión, y recientemente se han llegado a resultados del 60% de precisión (Hong et al, 2020). Además, se ha evidenciado que no siempre se tiene una buena correlación entre los indicadores de zona de flujo propuestos por Amaefule et al. (1993) y los registros eléctricos, ya que por ejemplo, en el registro neutrón se pueden tener arenas limpias de grano grueso de buena selección y alta porosidad, teniendo un indicador de zona de flujo alto al igual que la respuesta del registro, sin embargo, se puede tener así mismo respuestas altas del registro en litologías de grano fino donde el indicador de zona de flujo es pequeño, siendo la porosidad efectiva casi nula para zonas arcillosas. En el caso del registro de densidad, en un intervalo de arena limpia con muy alta permeabilidad donde el indicador de zona de flujo es alto se leen bajas densidades, y también se podrán leer bajas densidades para

intervalos con bajos indicadores de zona de flujo donde la roca está compuesta por granos muy finos y presenta alta porosidad, presencia de minerales livianos o espacio poroso saturado de gas (Fazel, 2014). Sin embargo, estos algoritmos han demostrado ser de gran utilidad en la identificación de electrofacies, por lo que en este paradigma de aprendizaje el entrenamiento se basará en el agrupamiento o clustering de ciertas combinaciones de respuestas de registros, posterior al respectivo análisis estadístico y filtrado de información, buscando la mejor precisión posible.

2.2.3 Unidades de Flujo

Amaefule et al. (1993) expuso una metodología de predicción de permeabilidad en pozos no corazonados a partir de modelos de regresión entre los registros eléctricos y los indicadores de zona de flujo (FZI). Define así, una unidad de flujo como un intervalo o subdivisión de roca con capacidades de flujo y almacenamiento similares, distintas a los demás intervalos. De esta forma, se establece que estas capacidades dependen de atributos geológicos texturales, mineralógicos, presencia de estructuras sedimentarias, contactos de fluidos y barreras de permeabilidad, ya que estos influyen en la porosidad efectiva y permeabilidad absoluta. Se reconoce la necesidad del correcto modelamiento de la permeabilidad para la planeación e implementación de estrategias de completamiento para el éxito de programas de recuperación, así como la construcción de modelos representativos de simulación. Por esto, si se tiene la suficiente heterogeneidad en el yacimiento, el modelo convencional de relación porosidad efectiva-permeabilidad no podrá implementarse para la determinación de la permeabilidad a partir de la porosidad efectiva, reflejando la presencia de más de un intervalo en el yacimiento con distinta correlación de estas propiedades, llamados

intervalos de flujo. A partir del estudio de la influencia de variables geológicas (depositacionales y diagenéticas) que controlan el flujo, Amaefule et al. (1993) propusieron una variable denominada indicador de zona de flujo (FZI), la cual está en función del índice de calidad de roca (RQI) y la porosidad efectiva normalizada (Øz), que están así mismo en función de la permeabilidad al aire (k) y la porosidad efectiva ($\emptyset_e$) de la roca. Así mismo, establecen una relación entre el indicador de zona de flujo y el área superficial por unidad de volumen de sólido (Sgv), la tortuosidad de la roca (τ) y un factor de forma (Fs) a partir de modificaciones en la ecuación de Kozeny-Carman de la siguiente forma:

$$RQI = 0.0314\sqrt{\frac{k}{\emptyset_e}} \tag{1}$$

$$\emptyset_z = \frac{\emptyset_e}{1-\emptyset_e} \tag{2}$$

$$FZI = \frac{1}{\sqrt{F_s}\tau S_{gv}} = \frac{RQI}{\emptyset_z} \tag{3}$$

Donde k se expresa en mD (milidarcies), RQI en μm (micrómetros), FZI en μm, $\emptyset_e$ y $\emptyset_z$ en fracción, F_s y τ adimensionales.

De las expresiones anteriores, obtuvieron lo siguiente

$$Log(RQI) = Log(\emptyset_z) + Log(FZI) \tag{4}$$

De manera que en un gráfico log-log de RQI vs $\emptyset_z$, todas las muestras con valores similares de FZI caerán en una recta con igual pendiente. Muestras con valores diferentes de FZI caerán sobre rectas paralelas entre sí. Y cada valor de FZI para cada grupo de muestras se puede obtener del intercepto de la recta de igual pendiente unitaria cuando $\emptyset_z = 1$. Siguiendo con la analogía planteada anteriormente, las muestras que caigan sobre la mima recta de pendiente unitaria tendrán atributos similares de sus gargantas de poro, y por ende constituyen una unidad de flujo. El indicador de zona de flujo se establece como un parámetro único que incorpora atributos geológicos de textura y mineralogía en la discriminación de distintas facies porales geométricas o unidades hidráulicas. En general, arenas con granos finos pobremente seleccionados y asociados a la presencia de minerales arcillosos tienden a exhibir una alta área superficial y alta tortuosidad, por ende, un bajo FZI. En contraste, arenas limpias de grano grueso y bien seleccionadas exhiben áreas superficiales menores, factores de forma menores, menor tortuosidad y por ende mayores valores de FZI. Igualmente, diferentes ambientes depositacionales y procesos diagenéticos controlan la geometría del poro y los FZI en rocas carbonatadas.

A pesar de obtener analíticamente el valor de los FZI para un conjunto de datos, la determinación del número exacto de unidades de flujo conlleva la aplicación de técnicas estadísticas convencionales como el uso de diagramas de frecuencia e histogramas, pruebas de normalidad, análisis de *clusters*, y pruebas de ensayo y error. Un histograma de frecuencias unidimensional del FZI acoplado con una prueba de normalidad convencional es usada para distinguir las familias de las unidades hidráulicas. Tal como es documentado por Amaefule et al. (2013), variables distribuidas unimodal e intrínsicamente en cualquier población homogénea son distribuidas aproximadamente de forma normal. Un gráfico de probabilidad de una variable unimodal y normalmente distribuida a menudo resulta en una línea recta. En contraste, la existencia

de múltiples subgrupos homogéneos en una población dada frecuentemente resulta en una distribución multimodal y por ende en múltiples líneas rectas en el gráfico de probabilidad. La identificación de las unidades hidráulicas fue llevada a cabo con el uso de la técnica de la media cuadrática, en donde se calculan los errores relativos para la determinación de la incertidumbre asociada al cálculo del FZI para cada recta de tendencia.

El proceso de caracterización por unidades de flujo se propone como parte de un proceso integral de caracterización de la roca yacimiento en donde se usan otras tecnologías y estudios como los análisis petrográficos (XRD, SEM, mineralogía FTIR) para determinar las características mineralógicas y texturales de la roca, pruebas de sensibilidad de estrés al índice de calidad de roca y presiones capilares para la caracterización de la garganta poral. De esta forma, la identificación del número de unidades de flujo hidráulicas se encontrará siempre determinado por los parámetros depositacionales y la influencia de procesos diagenéticos.

3. Análisis Exploratorio de Datos

Previo a la aplicación de cualquier técnica de aprendizaje automático e inteligencia artificial, se debe realizar un análisis exploratorio de datos cuyo objetivo es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas. Para esto, el análisis exploratorio de datos proporciona métodos sistemáticos sencillos para organizar y preparar los datos, detectar fallos en el diseño y adquisición de estos, tratar y evaluar los datos ausentes, identificar casos atípicos y comprobar los supuestos subyacentes en la mayoría de las técnicas multivariables. De esta forma, se combinan técnicas analíticas, estadísticas y gráficas para

proceder a realizar el preprocesamiento de los datos, el cual es el primer paso de cualquier modelo de *Machine Learning*. En el presente capítulo se expone el desarrollo del análisis de los datos, partiendo de la recopilación de data disponible, identificando las formaciones geológicas óptimas para el desarrollo de la metodología planteada, seguido de un recuento de registros, curvas y núcleos disponibles, y culminando con el respectivo análisis por formación geológica. Dado que no existe un flujo de trabajo estándar y generalizado para llevar a cabo el análisis exploratorio de datos, sino que depende del tipo de datos y situación a estudiar, se busca para el presente caso de estudio analizar las variables petrofísicas a modelar bajo un carácter interpretativo para garantizar consistencia en los resultados.

3.1 Data disponible

El conjunto de datos usados en la presente investigación se encuentra publicada en la web para acceso gratuito*, en patrocinio por la Agencia Nacional del Petróleo, Servicio Geológico de Brasil y el Ministerio de Minas y Energía de Brasil. Sin embargo, se evidenció que muchas zonas prospectivas de la cuenca no cuentan con un *set* de data completo o adecuado para su estudio, de manera que el número de análisis de núcleos y registros de los pozos es limitado. Además, no se cuenta con datos de producción histórica de intervalos específicos por pozo.

Dado que la metodología planteada implica el uso de porosidades efectivas y permeabilidades al aire obtenidas de análisis de núcleos, este fue el primer conjunto de datos a recopilar, para luego verificar la presencia de registros de pozo en dichos pozos y asegurar la data mínima requerida para el estudio.

* https://reate.cprm.gov.br/anp/TERRESTRE

La tabla 3 muestra los pozos que tienen análisis de núcleos rutinarios con su respectiva formación. Según los reportes encontrados en la página web consultada, la porosidad del núcleo fue obtenida mediante el método del porosímetro de Helio y la permeabilidad reportada consiste en la permeabilidad al aire. La permeabilidad Klinkenberg, la cual consiste en el valor corregido por el deslizamiento del gas en la medición de la permeabilidad al aire, no fue reportada en la data disponible para acceso gratuito en la página web, ni tampoco los parámetros de medición usados en laboratorio, por lo que en la presente investigación solo será posible trabajar con esta propiedad en términos de permeabilidad. En el caso de la porosidad medida, en el capítulo 4 se evaluará como efectiva o total según los modelos de porosidad obtenidos de los registros.

En la presente investigación, se tomaron los datos de permeabilidad al aire en la dirección horizontal dado que el grado de inclinación de los pozos no supera los 10° y se considera esta dirección como la predominante al flujo. Se identificó así mismo que estos pozos poseen *set* de registros básicos. Todos los datos de profundidad reportados en el presente trabajo están en profundidad medida (MD) y metros.

Tabla 3

Resumen de pozos con análisis de núcleos disponible según su formación geológica

Pozo	**Formación**	**Campo**
7-MGP-40D-BA	Pojuca	Miranga
7-RO-14-BA	Agua Grande	Remanso
7-JND-13D-BA	Agua Grande	Tangará - Jandaia
7-JND-3D-BA	Agua Grande	Tangará - Jandaia
7-CX-84D-BA	Candeias	Cexis

1-ALV-5-BA	Candeias	REC-T-197
7-RO-14-BA	Sergi	Remanso
7-BA-405D-BA	Sergi	Buracica
3-BRSA-1177D-BA	Sergi	Taquipe
6-BRSA-1225D-BA	Sergi	Araçás

Nota. De la tabla se puede observar que sólo el pozo 7-RO-14-BA tiene núcleos en más de una formación, en este caso, Agua Grande y Sergi, y que los pozos 7-JND-13D-BA y 7-JND-3D-BA son los únicos que se encuentran en el mismo campo. Elaboración propia.

Al observar los nombres de las formaciones presentes en la data disponible de la tabla 3, y según lo plasmado en la sección 2.2.1 de la descripción del sistema petrolífero de la cuenca, se cuenta con 3 formaciones popularmente conocidas por ser buenas rocas reservorio en la cuenca: Agua Grande, Sergi y Pojuca, siendo las dos primeras areniscas eólicas/fluviales y la última una arenisca deltaica. Sin embargo, dado que Candeias representa la roca generadora por excelencia en la cuenca, su caracterización implica métodos adicionales que sobrepasan el alcance y la técnica planteada en la presente investigación, por lo que la data de esta formación no será tenida en cuenta en los desarrollos posteriores.

De acuerdo con la apreciación anterior, se proceden a recopilar los registros disponibles para los pozos de las formaciones geológicas seleccionadas. En la tabla 4 se exponen los registros por formación encontrados, en donde se observa que se cuenta con un total de 139 registros, los cuales provienen de 76 pozos, lo que significa que se tienen varios pozos con registros en más de una de las tres formaciones presentes.

En la figura 6 se ilustra el número de pozos con registros según la formación evaluada en similitud, en donde se puede observar la cantidad de pozos que tienen registros en una única formación, en dos, y en tres.

Tabla 4

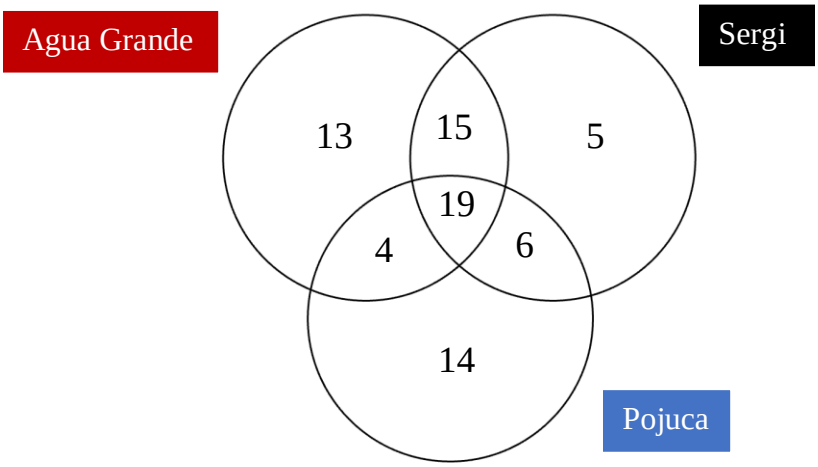
Resumen de registros por formación

Formación	**Pozos con registros**
Pojuca	43
Agua Grande	51
Sergi	45

Nota. De la tabla se puede apreciar que la formación Agua Grande representa la formación con mayor cantidad de pozos con registros. Así mismo, se identificó que todos los pozos con núcleos identificados en la tabla 1 tienen registros en los intervalos de las formaciones geológicas corazonadas.

Figura 6

Diagrama de Venn representando número de pozos con registros según su formación



Nota. Elaboración propia

Dado que en la presente investigación se busca correlacionar los resultados de la metodología de Amaefule en la determinación de unidades de flujo con el comportamiento de los registros de pozo mediante *machine learning*, los pozos seleccionados deben tener ambos conjuntos de datos disponibles. De la página web consultada se comprueba que todos los pozos con análisis de núcleos de la tabla 3 poseen registros. La tabla 5 resume el recuento de los registros disponibles por pozo.

Tabla 5

Recuento de curvas por pozo

Pozo / Registro	**GR**	**BS**	**CALI**	**RDEEP**	**RHOB**	**NPHI**	**PEF**	**DT**
7-MGP-40D-BA	X	X	X	X	X	X	X	X
7-RO-14-BA	X		X	X	X	X		
7-JND-3D-BA	X	X	X	X	X	X	X	X
7-JND-13D-BA	X	X	X	X	X	X	X	X
7-BA-405D-BA	X	X	X	X	X	X		
3-BRSA-1177D-B	X	X	X	X	X	X	X	X
6-BRSA-1225D-BA	X	X	X	X	X	X	X	X

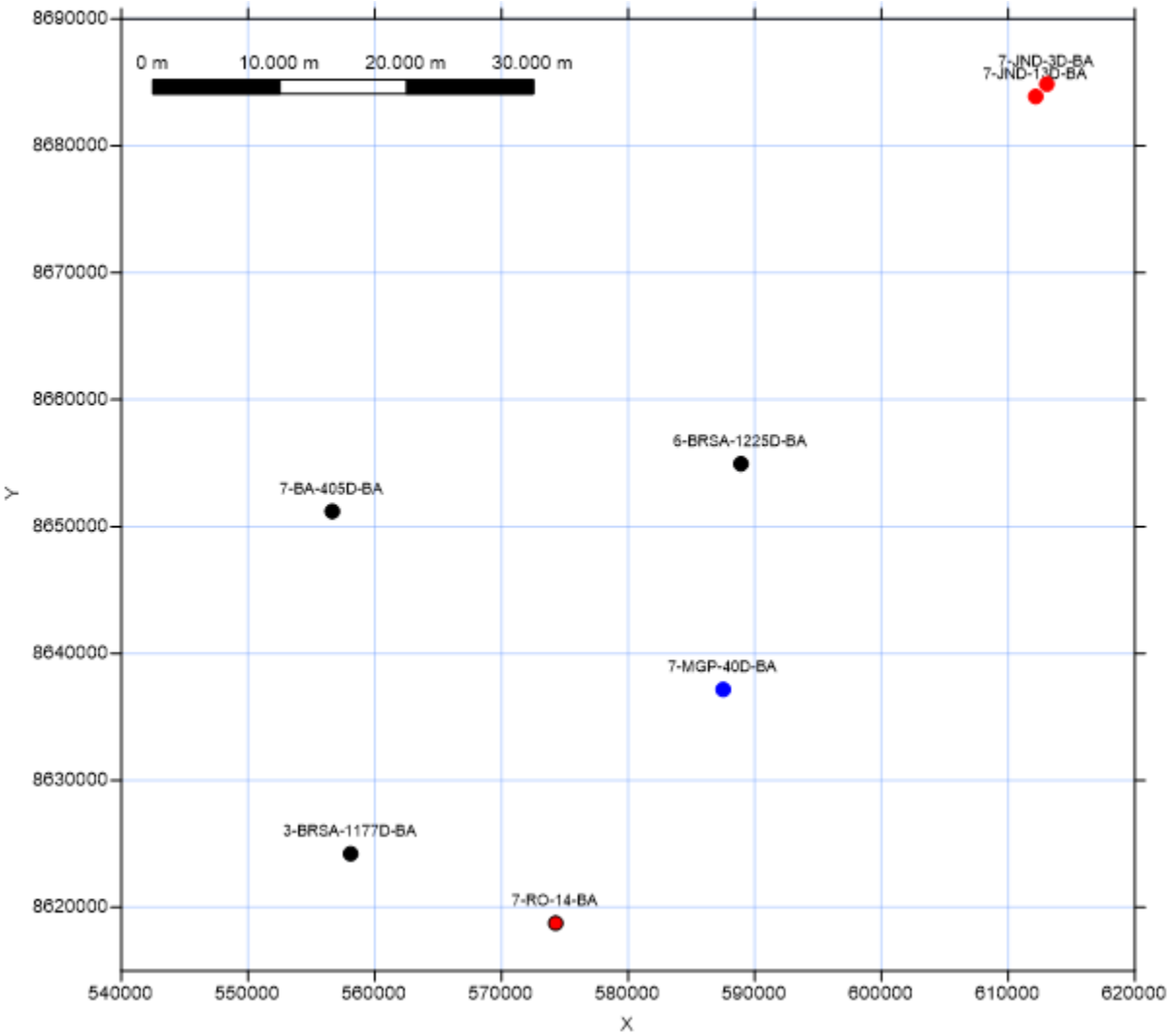
Nota. En la tabla se resumen las curvas presentes en los pozos escogidos para realizar el análisis. Elaboración propia

De la tabla 5, se puede observar que los pozos seleccionados no tienen las mismas curvas entre sí. En la preparación del conjunto de datos que será usado en los modelos de entrenamiento, se considera un primer factor de selección que consiste en la calidad de los registros disponibles de los pozos. Este se compone de la evaluación cualitativa de las curvas, es decir, comportamientos normales y congruentes eliminando afectaciones por condiciones del hoyo o errores de procesamiento, y del número de curvas disponible. En primera instancia, se observa que los pozos

7-RO-14-BA y *7-BA-405D-BA* son los que menos curvas tienen, pudiendo afectar el desarrollo de los modelos. Estos pozos podrían ser representativos según la cantidad de datos de análisis de núcleos disponible. Finalmente, en la figura 7 se observa la ubicación de los pozos seleccionados para el análisis.

Figura 7

Mapa de pozos seleccionados para el análisis



Nota. En el mapa se muestran los pozos seleccionados para el análisis de datos y la técnica propuesta. De color azul los pertenecientes a la formación Pojuca, de color negro a la formación

Sergi, de color rojo a la formación Agua Grande y en color rojo con borde negro a estas dos últimas. Elaboración propia.

3.2 Fm Pojuca

Como se mostró en la sección anterior, para la formación Pojuca solo se tiene un pozo con información de núcleos y registros, el ***7-MGP-40D-BA***, perteneciente al campo Miranga, uno de los más importantes de la cuenca Recóncavo. De manera inicial, se procede a correlacionar los datos de núcleos con el registro *coregamma* para colocar la data a profundidad, realizando un desplazamiento de +8 metros de profundidad para dicho propósito. El pozo cuenta con 194 datos de porosidad de Helio y 194 datos de permeabilidad al aire.

El siguiente paso consiste en evaluar el comportamiento estadístico de los datos de porosidad y permeabilidad al aire a través del uso de histogramas de frecuencia, gráficos de distribución de probabilidad y cálculos de medidas de tendencia central.

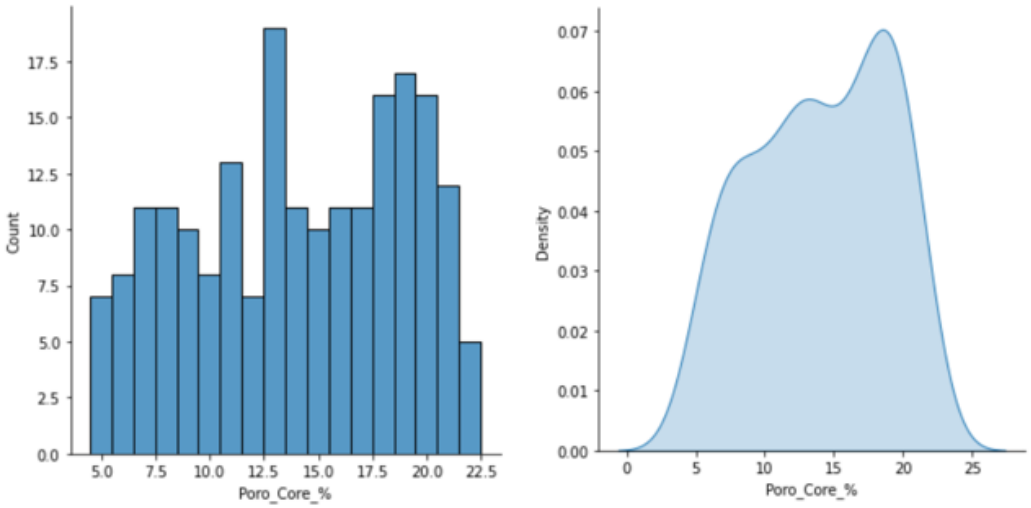
La figura 8 muestra un comportamiento trimodal para la porosidad en el intervalo corazonado de 113 metros para este pozo. Este comportamiento característico refleja la presencia de más de una condición de distribución de tamaño poroso en el intervalo, representando varios litotipos con posibles variaciones de litología o mineralogía. En el caso del logaritmo de la permeabilidad al aire en la figura 9, se observa un leve comportamiento bimodal, evidenciando la presencia de distintas condiciones de flujo y la influencia de más de una propiedad sobre el valor de la permeabilidad al compararse con el gráfico de distribución de porosidad de la figura 8.

Por último, se realiza el método gráfico Q-Q (cuantil-cuantil) para el diagnóstico de diferencias de la distribución de probabilidad de la muestra con la teórica normal. En la figura 10

se observa el resultado del método para la porosidad del núcleo y el logaritmo de la permeabilidad al aire.

Figura 8

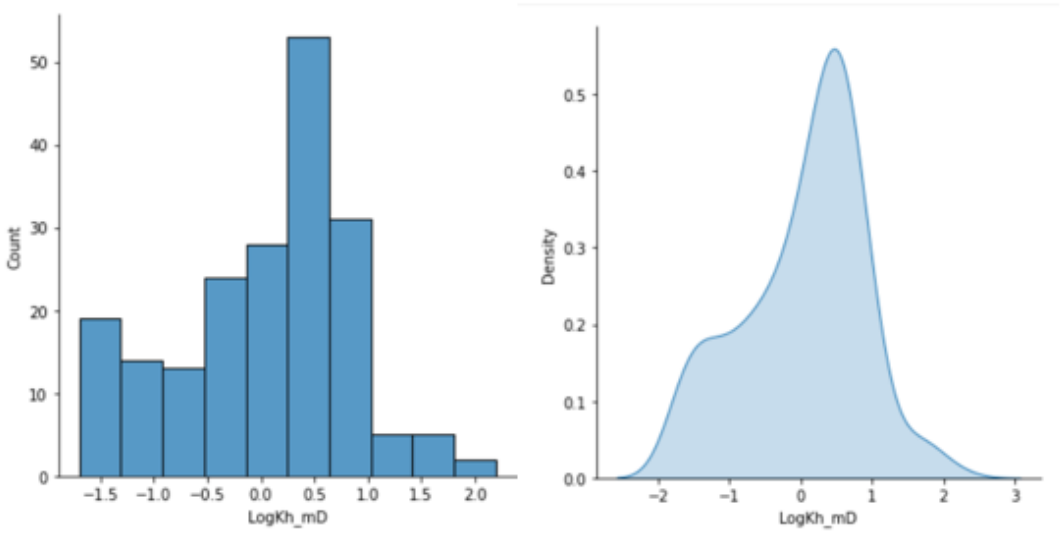
Análisis estadístico de porosidad del pozo 7-MGP40D-BA



Nota. Elaboración propia

Figura 9

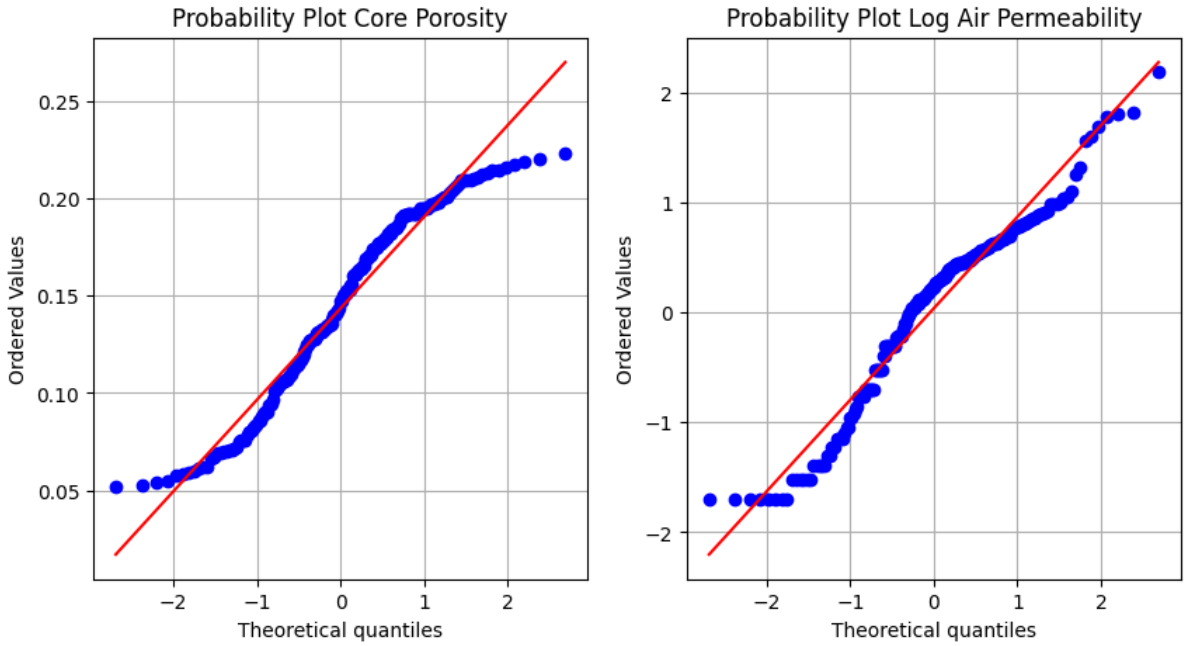
Análisis estadístico del logaritmo de la permeabilidad al aire del pozo 7-MGP-40D-BA



Nota. Elaboración propia

Figura 10

Gráfico Q-Q para la porosidad del núcleo y log. permeabilidad al aire en el pozo 7-MGP-40D-BA



Nota. Elaboración propia

De la figura 10 se confirma el comportamiento multimodal identificado en los histogramas de las figuras 8 y 9, en donde tanto los datos de porosidad del núcleo y el logaritmo de la permeabilidad al aire presentan concentración de datos en zonas no esperadas, alejándose la distribución normal teórica (línea roja). De esta forma, se confirma la presencia de distintas condiciones de flujo y almacenamiento en el intervalo corazonado.

De la tabla 6 según las medidas de tendencia central de la porosidad y permeabilidad al aire, se puede observar que, si bien los valores de porosidad se concentran alrededor de una mediana regular de 13%, los valores de permeabilidad tienen una mediana de 1.7 mD, cuyas condiciones de flujo no representan interés para ser caracterizadas como unidades de flujo hidráulicas bajo la metodología planteada.

Tabla 6

Medidas de tendencia central de las propiedades petrofísicas del pozo 7-MGP-40D-BA

Medida [m]	**Porosidad [frac]**	**Permeabilidad al aire [mD]**
Media	0.143	4.95
Desviación Est.	0.047	14.44
Mínimo	0.052	0.02
25%	0.106	0.30
Mediana	0.147	1.70
75%	0.185	4.05
Máximo	0.223	156.6

Nota. Elaboración propia

3.3 Fm Agua Grande

Para el caso de la formación Agua Grande, se cuenta con los pozos ***7-RO-14-BA***, ***7-JND-3D-BA*** y ***7-JND-13D-BA***. Al momento de recolectar la data disponible, se encontró que el pozo ***7-RO-14-BA*** no contaba con suficiente data de análisis de núcleos para realizar el análisis interpretativo del comportamiento estadístico de las propiedades petrofísicas de la formación Agua Grande, lo que también impide correlacionar la porosidad y el logaritmo de la permeabilidad al aire y tener un conjunto de datos robusto para modelar el indicador de zona de flujo y obtener resultados consistentes en el modelo de aprendizaje, por lo que se descarta su análisis.

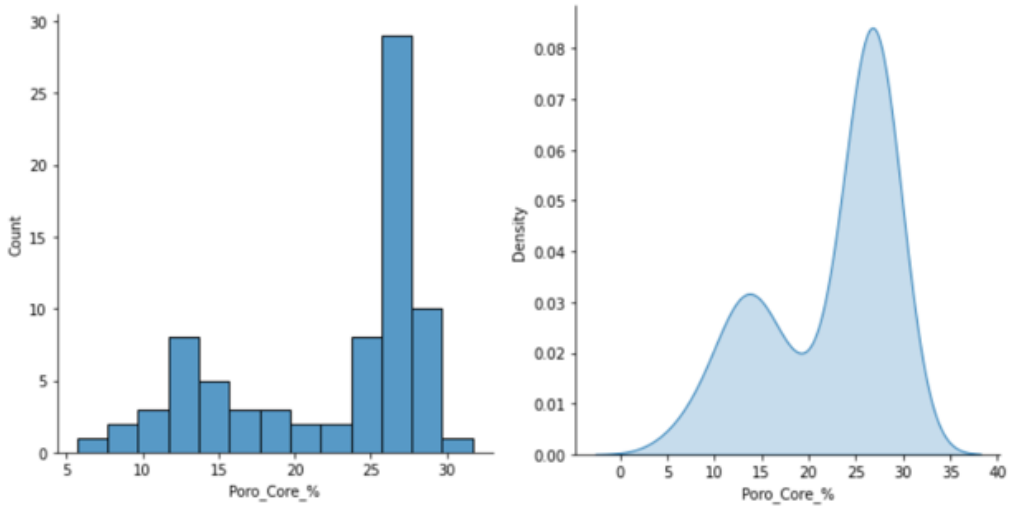
En las figuras 11 y 12 se muestra el análisis de la porosidad y el logaritmo de la permeabilidad al aire para el primer pozo de Agua Grande en estudio, el ***7-JND-3D-BA***. El pozo cuenta con 77 datos de porosidad de Helio y 77 datos de permeabilidad. Para este pozo se observa

un posible comportamiento bimodal en la distribución de probabilidad de la porosidad del núcleo y logaritmo de la permeabilidad al aire. En un caso similar al estudiado para el pozo de la formación Pojuca, se refleja la presencia de distintos litotipos en el intervalo corazonado con distintas distribuciones de tamaños de poro que generan estos subgrupos o modas en los gráficos. Esta condición puede representar la presencia de distintas unidades de flujo. En la figura 13 se observa el gráfico Q-Q para la porosidad del núcleo y el logaritmo de la permeabilidad al aire. Para este caso, se observa de la figura 13 un comportamiento bimodal claro en ambas variables al marcar dos tendencias mucho más planas que la distribución teórica normal.

De la tabla 7 según las medidas de tendencia central de la porosidad y permeabilidad al aire, se evidencian muy buenos valores para ambas propiedades, teniendo la permeabilidad una mediana mayor a 1000 mD, alcanzando máximos de 3770 mD sin valores atípicos. Esta condición hace atractivo el análisis de unidades de flujo para el pozo bajo la metodología planteada.

Figura 11

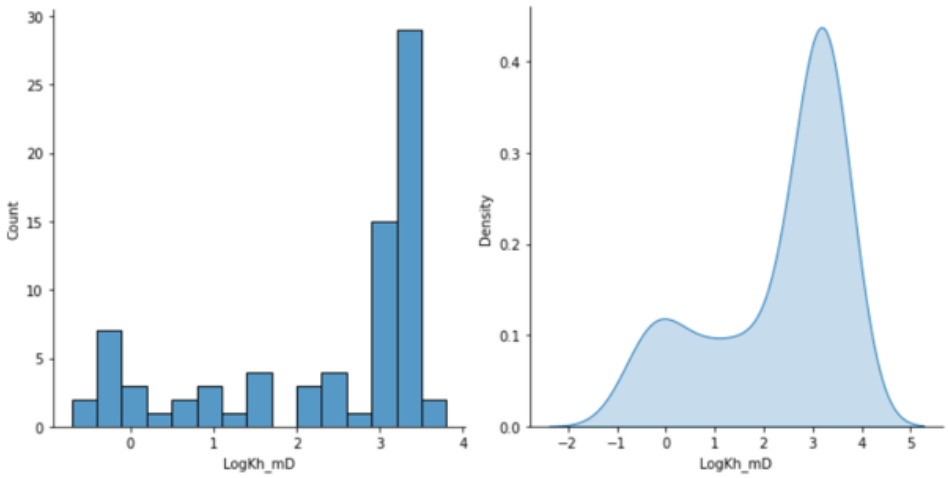
Análisis estadístico de porosidad de núcleo del pozo 7-JND-3D-BA



Nota. Elaboración propia

Figura 12

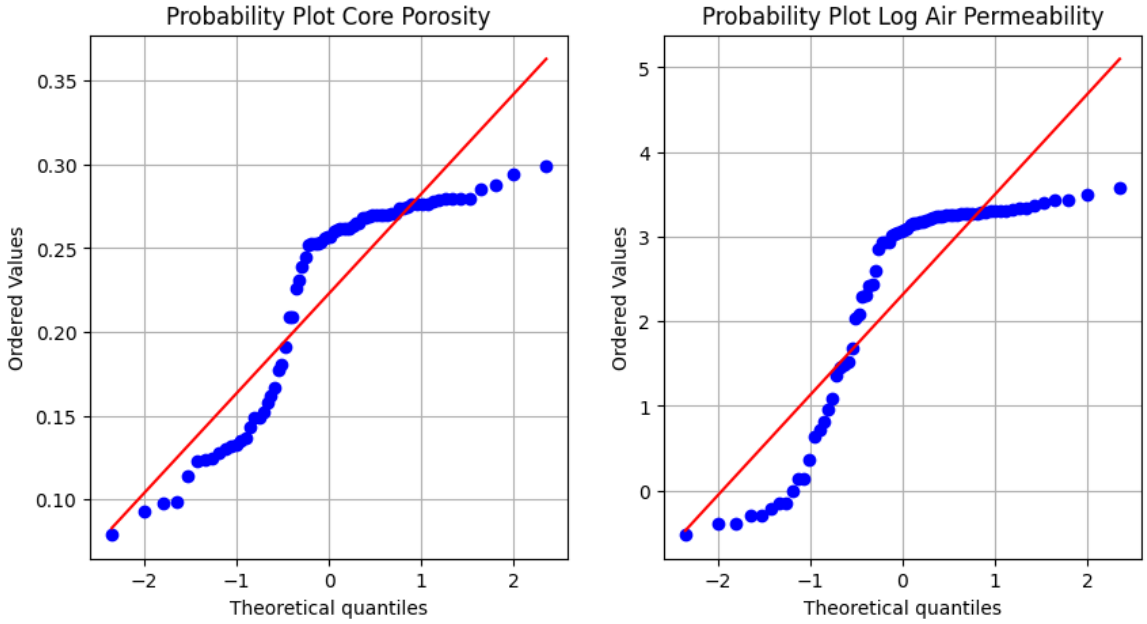
Análisis estadístico del logaritmo de la permeabilidad al aire del pozo 7-JND-3D-BA



Nota. Elaboración propia

Figura 13

Gráfico Q-Q para la porosidad del núcleo y log. permeabilidad al aire en el pozo 7-JND-3D-BA



Nota. Elaboración propia

Tabla 7

Medidas de tendencia central de las propiedades petrofísicas del pozo 7-JND-3D-BA

Medida [m]	**Porosidad [frac]**	**Permeabilidad al aire [mD]**
Media	0.222	1108.2
Desviación Est.	0.066	964.1
Mínimo	0.057	0.2
25%	0.158	28.9
Mediana	0.257	1199.5
75%	0.271	1879.6
Máximo	0.299	3770.8

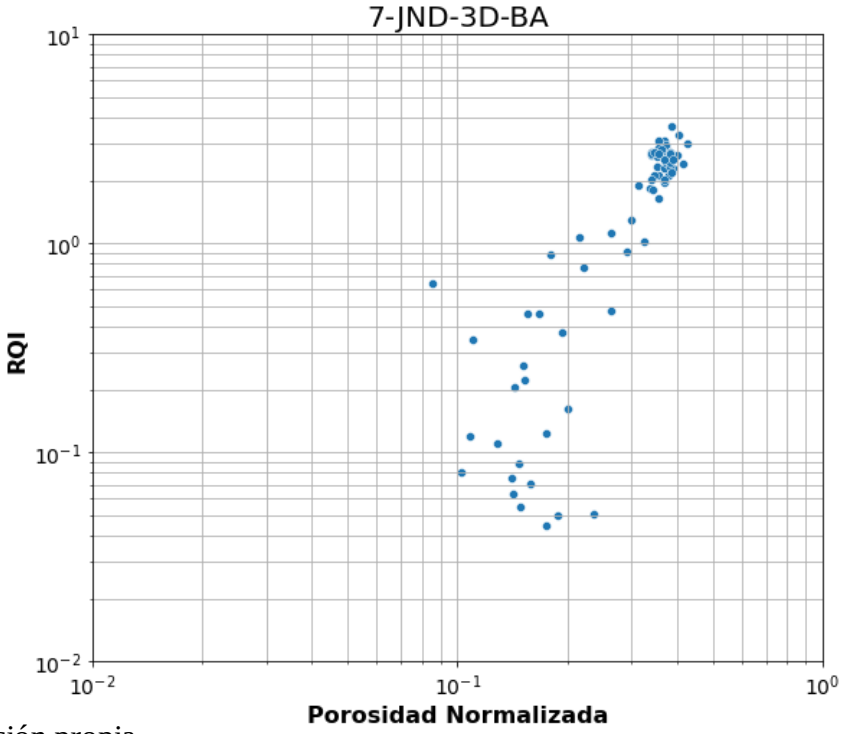
Nota. Elaboración propia

Siguiendo los pasos comentados en la sección 2.2.3 de la metodología de Amaefule, se calcula el RQI y porosidad normalizada a partir de los datos de porosidad medida y permeabilidad al aire y con estos el FZI. En la figura 14 se muestra la relación entre el RQI y la porosidad normalizada calculados en escala logarítmica.

De la figura 14 se pueden observar dos grandes tendencias en la distribución de los datos. Primeramente, se observa una alta concentración de datos en un rango de RQI de 1 a 3 y una porosidad normalizada de 0.3 a 0.43. La segunda tendencia se muestra como una alta dispersión en los datos por debajo de la primera tendencia. De acuerdo con lo planteado por Amaefule, esta condición representa la existencia de más de una unidad de flujo hidráulica en el intervalo analizado.

Figura 14

Relación entre RQI y Porosidad normalizada para el pozo 7-JND-3D-BA



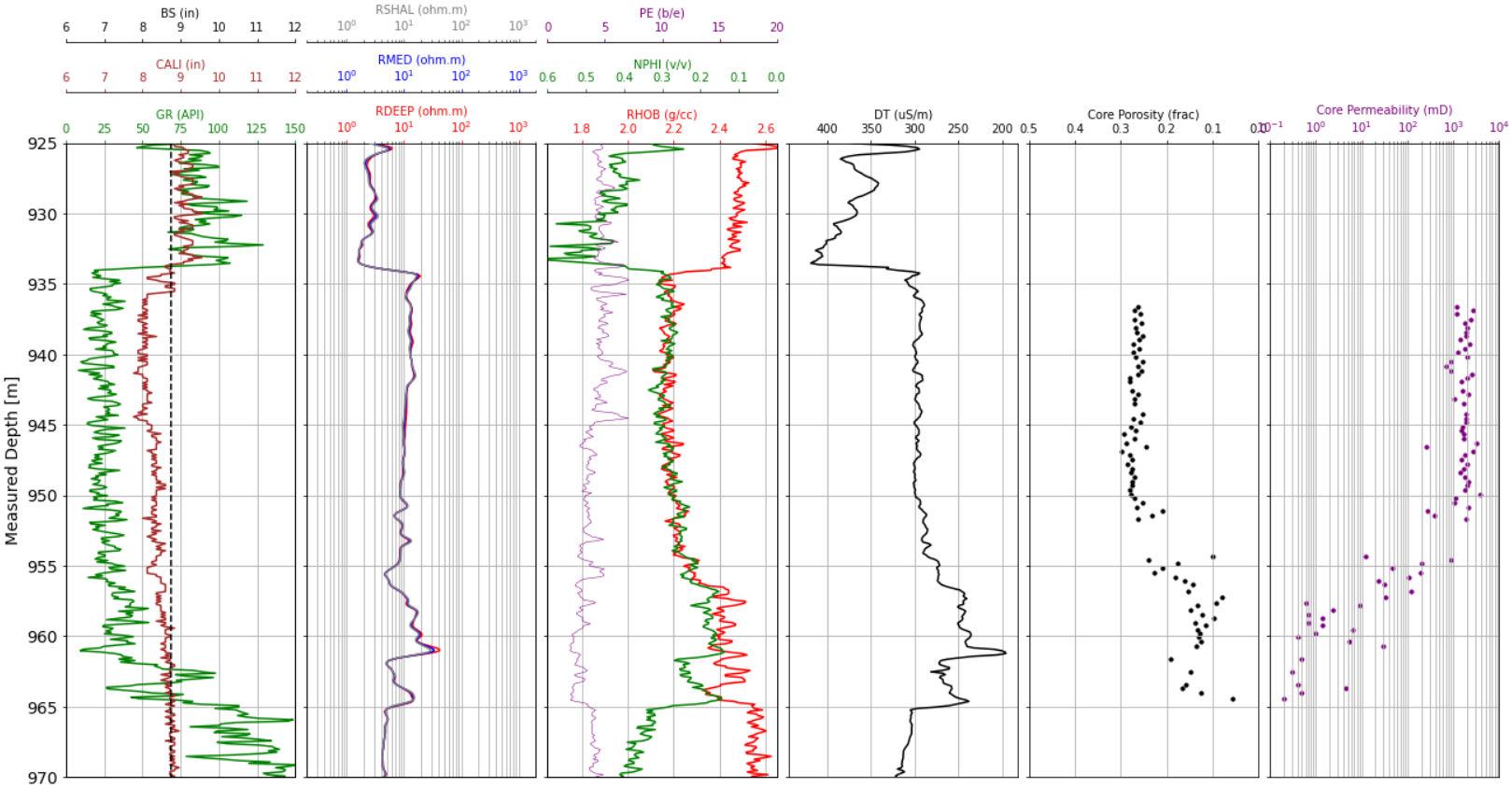
Nota. Elaboración propia

Dado que el pozo no tiene registro de *coregamma* disponible para colocar los datos de núcleo a profundidad, se realizó la correlación con la data de porosidad del núcleo y los registros de porosidad del pozo, obteniendo un desplazamiento de +2 metros de profundidad. Por último, en la figura 15 se muestra el conjunto de registros y análisis de núcleos para el pozo ***7-JND-3D-BA***.

De la figura 15, se observa un cuerpo continuo de baja radioactividad según el comportamiento del GR (curva verde pista uno) a partir de 934 m hasta 964 m, con presencia de minerales radioactivos hacia la base del intervalo. Según la información disponible del campo Jandaia, este intervalo corresponde a la formación Agua Grande del presente pozo. Así mismo, se observan aumentos en la resistividad leída, con máximos cercanos a 40 Ohm.m.

Figura 15

Registros y análisis de núcleos para el pozo 7-JND-3D-BA



Nota. Elaboración propia usando el lenguaje Python. Profundidades reportadas en metros.

La tercera pista del *template* muestra la densidad y el neutrón en escala de arenisca y se puede apreciar la poca separación de las curvas en una gran parte de la formación, indicando la presencia de una arenisca limpia con porosidades superiores al 30%. Hacia la base del intervalo, se comienza a apreciar la separación de las curvas junto con un aumento del GR, indicando el aumento de la presencia de minerales arcillosos. El factor fotoeléctrico por otro lado muestra valores altos de 5 b/e, lo cual no corresponde con los valores reportados para las areniscas. En este sentido, la formación de un *mudcake* evidenciado por el caliper puede estar influenciando la medición del registro. Dado el ambiente sedimentario fluvial de la formación Agua Grande, es posible que durante el proceso de transporte o limpieza realizado por las corrientes de agua es

posible que parte del material presente no haya sido completamente lavado, teniendo posibles presencias de feldespatos, biotitas o moscovitas, que tienden a elevar el valor leído de la herramienta. Sin embargo, la determinación de modelos mineralógicos está fuera del alcance de la presente investigación y sus comentarios se plantean como futuras recomendaciones.

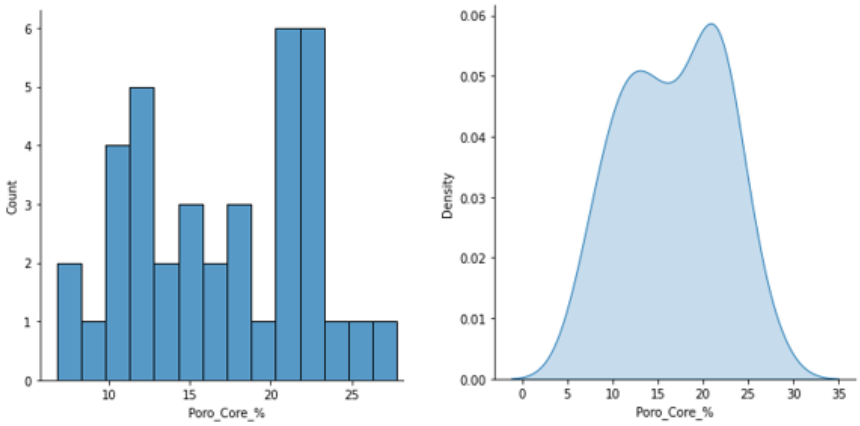
Por último, en la figura 15 se aprecian dos tendencias sobre los registros en la formación. La primera cubre la zona superior y media de la formación, dominada por poca arcillosidad, y la segunda la zona inferior de la formación, en donde se observa una reducción en los registros de porosidad y una mayor arcillosidad. Estas dos tendencias se relacionan con el comportamiento bimodal encontrado en las figuras 11 y 12.

En las figuras 16 y 17 se muestra el análisis de la porosidad del núcleo y el logaritmo de la permeabilidad al aire para el siguiente pozo de Agua Grande en estudio, el ***7-JND-13D-BA***. Para este pozo se cuenta con 38 datos de porosidad de Helio y 38 datos de permeabilidad al aire. Como se pueden observar de las figuras, las propiedades petrofísicas del pozo presentan un comportamiento bimodal al igual que el pozo anterior, los cuales como se estableció en la tabla 3 corresponden a pozos de un mismo campo.

En la figura 18 se muestra la prueba gráfica Q-Q para la porosidad y logaritmo de la permeabilidad al aire del pozo. Se observa primeramente un comportamiento más cercano a la distribución normal en la porosidad en comparación con el pozo anterior de Agua Grande, sin embargo, la presencia de las dos modas encontradas en el histograma de la figura 16 genera la dispersión observada. En el caso del logaritmo de la permeabilidad al aire, las dos modas del histograma de la figura 17 se observan como tendencias planas en el gráfico, una con mayor concentración de datos que otra. Estas condiciones representan la existencia de más de una condición de flujo y almacenamiento en el intervalo corazonado.

Figura 16

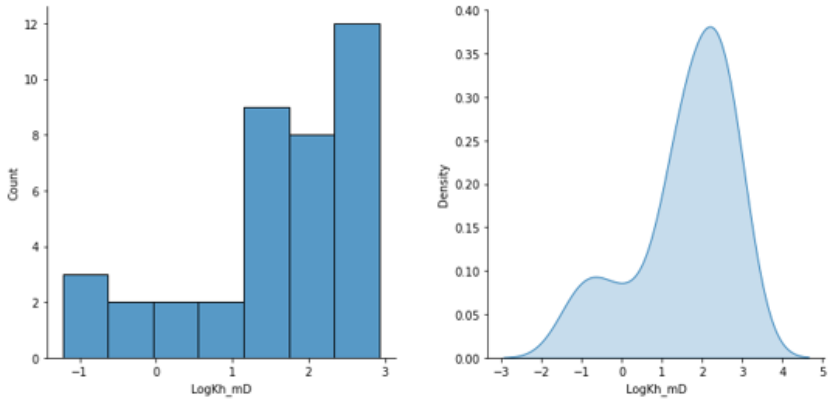
Análisis estadístico de porosidad de núcleo del pozo 7-JND-13D-BA



Nota. Elaboración propia

Figura 17

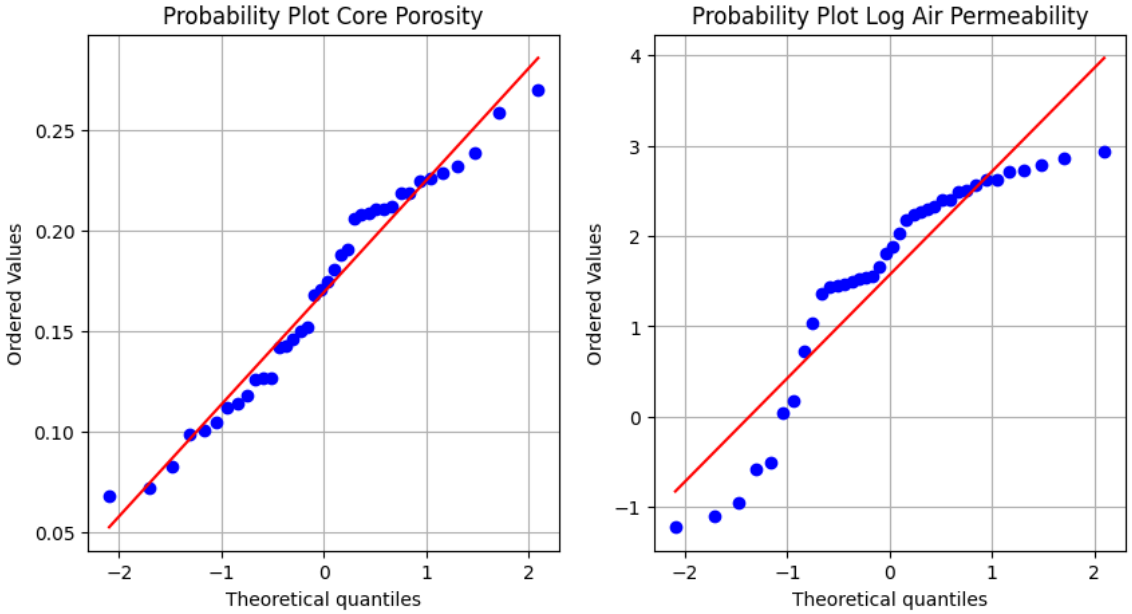
Análisis estadístico del logaritmo de la permeabilidad al aire del pozo 7-JND-13D-BA



Nota. Elaboración propia

Figura 18

Gráfico Q-Q para la porosidad del núcleo y log. permeabilidad al aire en el pozo 7-JND-13D-BA



Nota. Elaboración propia

De la tabla 8 según las medidas de tendencia central de la porosidad y permeabilidad al aire, se evidencian evidencia buenos valores para ambas propiedades, teniendo la permeabilidad una mediana de 70.5 mD y alcanzando máximos de 848 mD, con dos valores atípicos. Esta condición hace atractivo el análisis de unidades de flujo para el pozo bajo la metodología planteada.

En la figura 19 se muestra el gráfico logarítmico de RQI vs Porosidad Normalizada siguiendo los cálculos propuestos por Amaefule. De forma evidente, se observa una tendencia lineal importante en el gráfico, cubriendo una gran cantidad de datos. Además, una pequeña concentración de datos dispersos se observa en la parte inferior del gráfico. Para este caso, se podría tener la presencia dos o más unidades de flujo, según las consideraciones respectivas del análisis de *clusters* necesario para definir las unidades en el capítulo 4.

Tabla 8

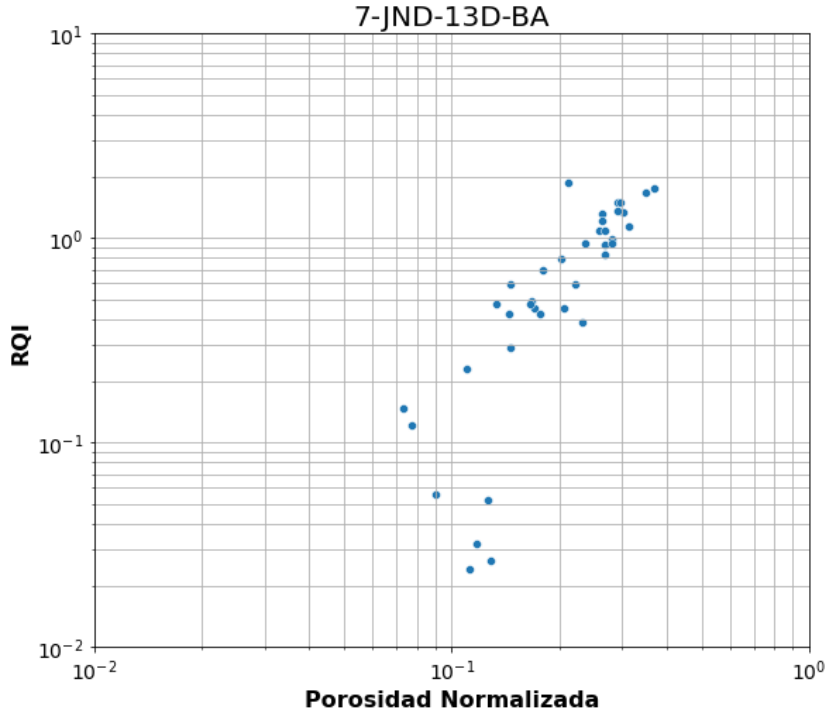
Medidas de tendencia central de las propiedades petrofísicas del pozo 7-JND-13D-BA

Medida [m]	**Porosidad [frac]**	**Permeabilidad al aire[mD]**
Media	0.169	185.2
Desviación Est.	0.054	224.9
Mínimo	0.068	0.06
25%	0.126	24
Mediana	0.173	70.5
75%	0.211	296
Máximo	0.270	848

Nota. Elaboración propia

Figura 19

Relación entre RQI y Porosidad normalizada para el pozo 7-JND-13D-BA



Nota. Elaboración propia

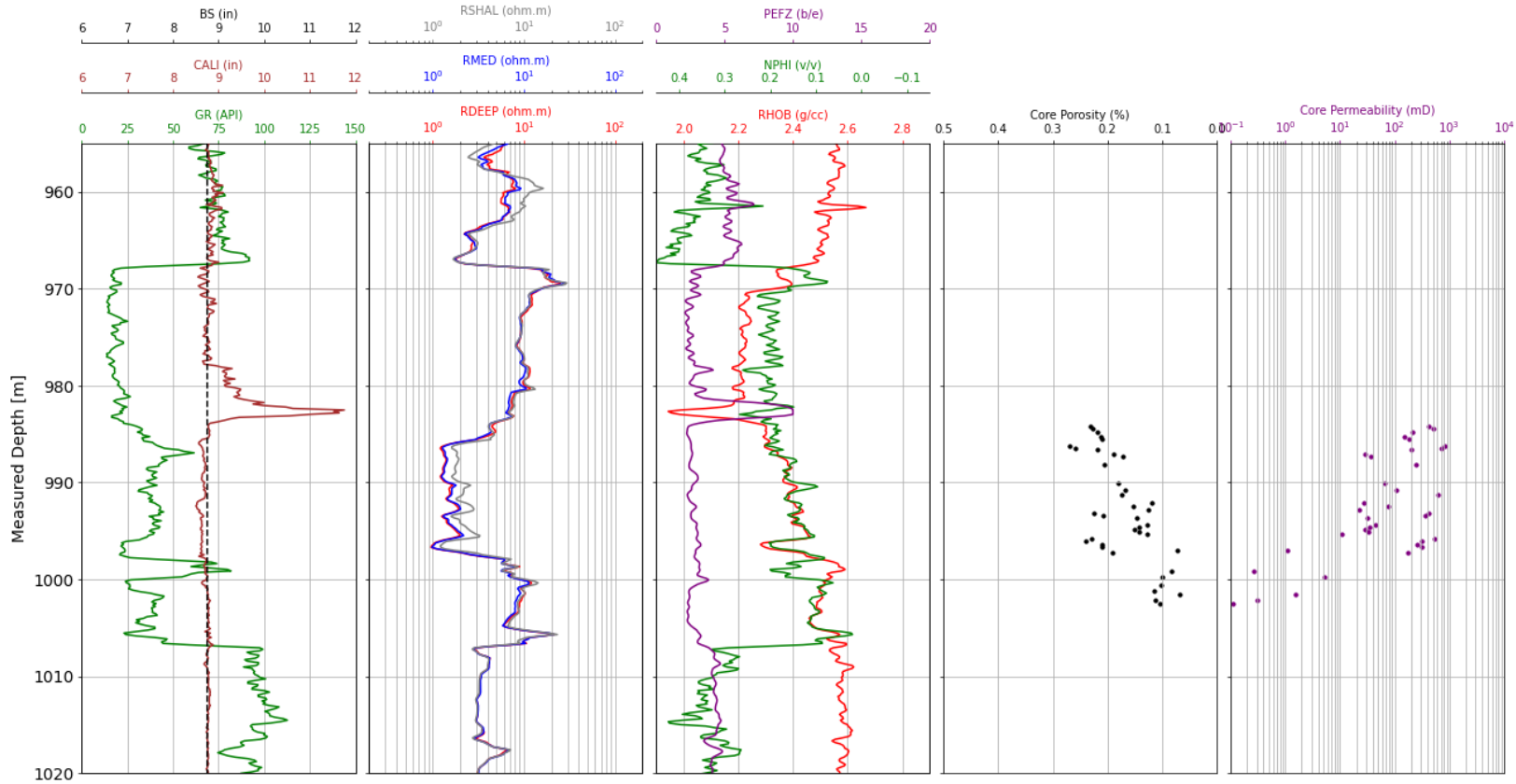
Dado que el pozo no tiene registro de *coregamma* disponible se procedió a realizar la calibración de las profundidades con los datos de porosidad del núcleo y registros de porosidad del pozo, obteniendo un desplazamiento de +2.5 m para los datos menores a 995.2 m y +4 m para los datos de una profundidad mayor a 995.2 m.

Por último, en la figura 20 se muestra el conjunto de registros y análisis de núcleos para el pozo ***7-JND-13D-BA***. Se puede observar en la parte superior del registro una afectación en las mediciones de las curvas de densidad y registro fotoeléctrico por mala condición del hoyo según lo evidenciado por el caliper, el cual muestra lecturas mayores a 9 pulgadas entre 982 metros y 983 metros. Estas circunstancias serán tenidas en cuenta en el desarrollo del algoritmo de aprendizaje filtrando los datos cuyas lecturas de caliper evidencien mala calidad del hoyo.

De la figura 20 se puede observar también un intervalo de baja radioactividad según el GR entre 967 m y 1006 m. De acuerdo con la información disponible del campo Jandaia este intervalo representa la formación Agua Grande, teniendo al igual que el pozo ***7-JND-3D-BA***, una zona de alta radioactividad hacia su base. Según el conjunto de datos disponible para ambos pozos, se evidencian dos diferencias principales. Primero, la resolución de los registros para el primer pozo ***7-JND-3D-BA*** (figura 15) es mayor que para el segundo pozo ***7-JND-13D-BA*** (figura 20). Segundo, los análisis de núcleos disponibles para el segundo pozo no abarcan el intervalo de la formación en su totalidad. Esta segunda diferencia genera que en el gráfico logarítmico del RQI vs Porosidad Normalizada en la figura 19, correspondiente al segundo pozo, no se alcance a apreciar la alta concentración de datos con mayores valores para estas propiedades que se observa para el primer pozo (figura 14). Los registros densidad y neutrón para este pozo se presentan por disponibilidad en matriz caliza, por lo que el *crossover* en la zona superior representa la presencia de areniscas y la poca separación inferior la arcillosidad.

Figura 20

Registros y análisis de núcleos para el pozo 7-JND-13D-BA

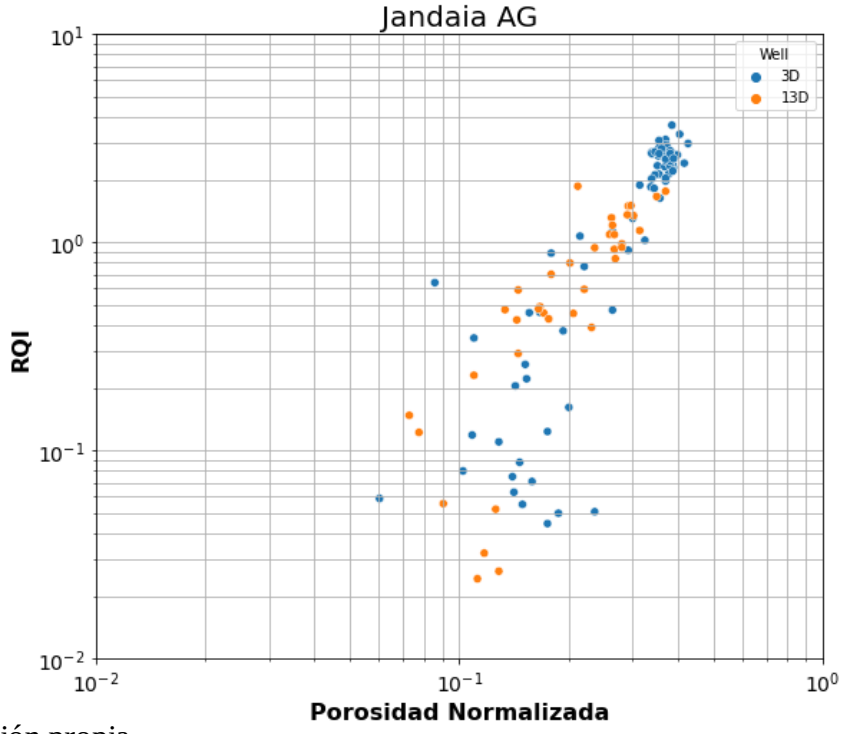


Nota. Elaboración propia usando el lenguaje Python. Profundidades reportadas en metros.

En la figura 21 se observa la relación entre RQI y la porosidad normalizada para los dos pozos analizados del campo Jandaia. El gráfico permite evidenciar la consistencia en el conjunto de datos para ambos pozos, mostrando tendencias similares para las propiedades calculadas. Esta situación representa la oportunidad de analizar la formación Agua Grande como un caso de estudio para los dos pozos del campo Jandaia con núcleo y registros. En donde se pueda evaluar el modelo de aprendizaje en ambos pozos, y así mismo, se deja a futura recomendación la calibración y validación del modelo en los pozos no corazonados del campo con análisis petrográficos y datos de productividad, no disponibles en la web.

Figura 21

Relación entre RQI y Porosidad normalizada para el campo Jandaia Fm Agua Grande



Nota. Elaboración propia

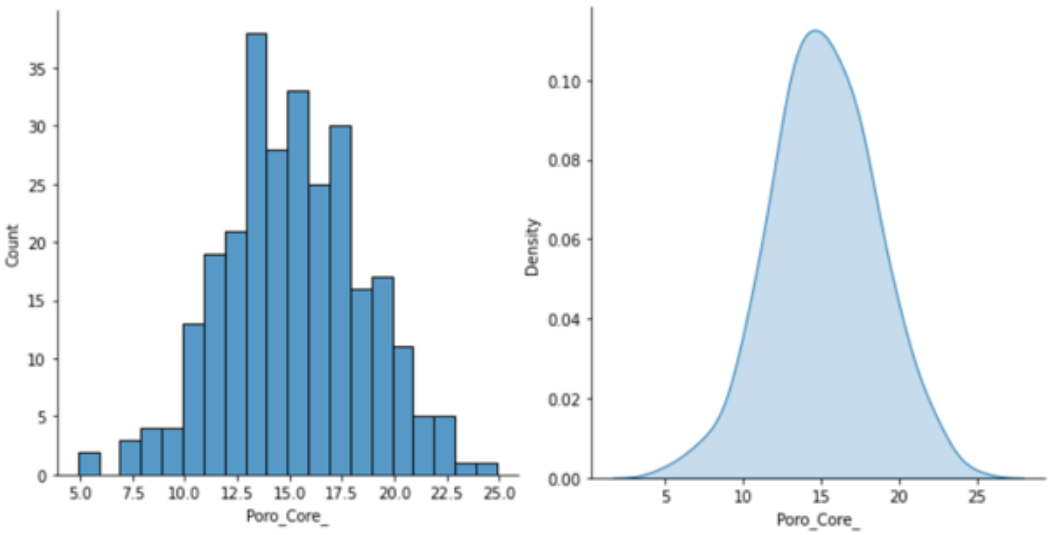
3.4 Fm Sergi

Para el caso de la formación Sergi, se cuenta con cuatro pozos para el análisis: ***7-RO-14-BA***, ***7-BA-405D-BA***, ***3-BRSA-1177D-BA*** y el ***6-BRSA-1225D-BA***. Al momento de recolectar la data disponible para estos pozos, se observó la poca data para los pozos ***7-BA-405D-BA***, ***3-BRSA-1177D-BA*** y el ***6-BRSA-1225D-BA***, lo cual impide su análisis ya que no se tiene la suficiente información para interpretar las posibles tendencias en la distribución de probabilidad de las propiedades petrofísicas de la formación Sergi en los pozos y la correlación entre la porosidad efectiva y la permeabilidad.

El pozo cuenta con 264 datos de porosidad de Helio y 264 datos de permeabilidad al aire. En las figuras 22 y 23 se muestra el análisis estadístico de los datos de núcleo para el pozo ***7-RO-14-BA***. En estas, se observa un comportamiento unimodal con distribución aproximadamente normal para la porosidad del núcleo y el logaritmo de la permeabilidad al aire. Sin embargo, se alcanza a apreciar una desviación estándar considerable afectando la forma del gráfico, sobre todo para el caso del logaritmo de la permeabilidad al aire en la figura 23. Este pozo, al igual que el pozo ***7-MGP-40D-BA*** analizado en la formación Pojuca, son los pozos con mayor análisis de núcleos disponibles, encontrando en ambos una alta desviación de los datos. Sin embargo, en el caso del presente pozo no se encuentran múltiples modas para la distribución de las variables. La alta desviación puede estar asociada al ambiente sedimentario fluvial/eólico de la formación Sergi, que al igual que el ambiente sedimentario deltaico de la formación Pojuca, se caracteriza por tener numerosos canales de arenas con diferentes capacidades de flujo y alta heterogeneidad vertical.

Figura 22

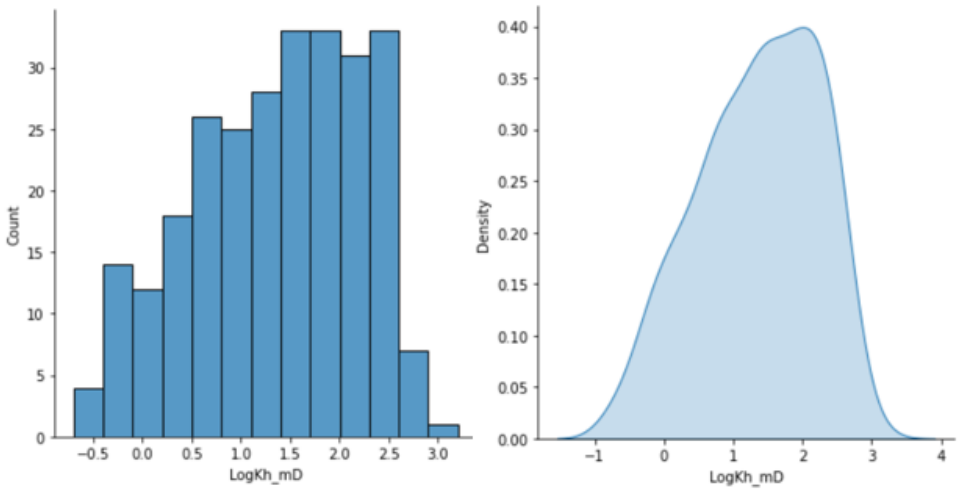
Análisis estadístico de porosidad de núcleo del pozo 7-RO-14-BA



Nota. Elaboración propia

Figura 23

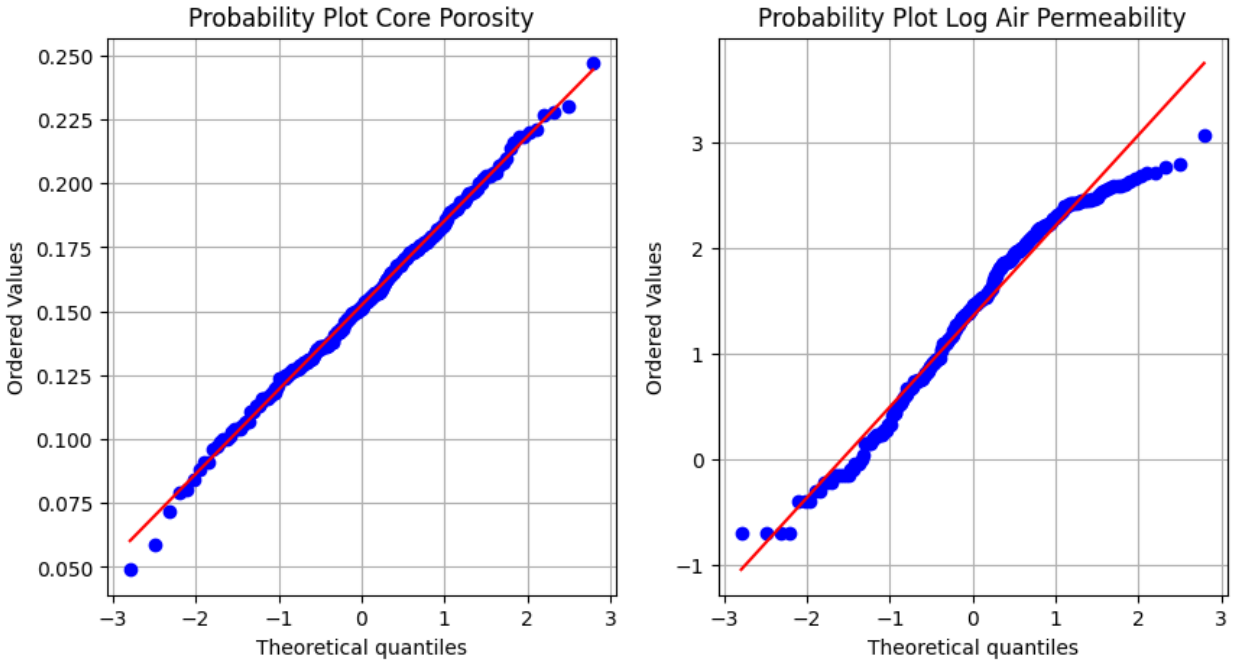
Análisis estadístico del logaritmo de la permeabilidad al aire del pozo 7-RO-14-BA



Nota. Elaboración propia

Figura 24

Gráfico Q-Q para la porosidad del núcleo y log. permeabilidad al aire en el pozo 7-RO-14-BA



Nota. Elaboración propia

En la figura 24 se observa la prueba gráfica Q-Q para la porosidad del núcleo y el logaritmo de la permeabilidad al aire del pozo ***7-RO-14-BA***. Se puede considerar que la porosidad del núcleo cuenta con una distribución normal unimodal aceptable, sin embargo, el logaritmo de la permeabilidad al aire presenta una desviación de la distribución normal en la parte alta del gráfico. Este tipo de desviación es conocida como *left skew* o *negative skew* (sesgo a la izquierda o negativo) la cual genera una distribución asimétrica a la izquierda, teniendo el gráfico de distribución de probabilidad una cola larga en su lado izquierdo, tal como se observa en la figura 23 para el logaritmo de la permeabilidad al aire. Dado que el modelo de *machine learning* (Modelos mixtos de Gauss) usado en el capítulo 4 para la determinación de las unidades de flujo no es considerablemente sensible al sesgo de los datos, no se considera la transformación del logaritmo de la permeabilidad al aire para reducir el seso negativo observado.

De la tabla 9 según las medidas de tendencia central de la porosidad y permeabilidad al aire, se pueden observar valores regulares para ambas propiedades, sobretodo en el caso de la permeabilidad al aire en donde la mediana es de 27.9 mD. Sin embargo, dada la alta cantidad de datos disponibles se considera el pozo como candidato a estudiar. Elaborando el gráfico logarítmico entre el RQI y porosidad normalizada calculados con la data del núcleo, figura 25, se encuentra una dispersión importante de los datos y la posible presencia de múltples tendencias, lo que llevaría a la presencia de más de una unidad de flujo hidráulica. Esta condición resulta favorable para la aplicación de la técnica planteada ya que se podrán correlacionar las múltiples tendencias encontradas con la respuesta de los registros disponibles para el pozo.

Tabla 9

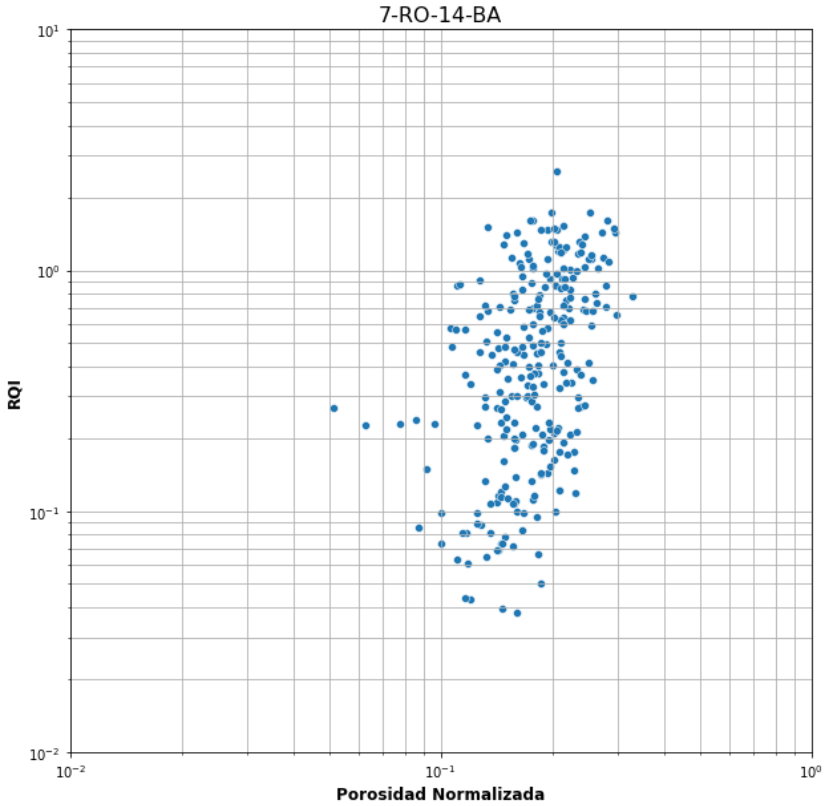
Medidas de tendencia central de las propiedades petrofísicas del pozo 7-RO-14-BA

Medida [m]	**Porosidad [frac]**	**Permeabilidad al aire [mD]**
Media	0.152	87.9
Desviación Est.	0.032	136.3
Mínimo	0.049	0.2
25%	0.13	5.4
Mediana	0.151	27.9
75%	0.174	117
Máximo	0.247	1157

Nota. Elaboración propia

Figura 25

Relación entre RQI y Porosidad normalizada para el pozo 7-RO-14-BA



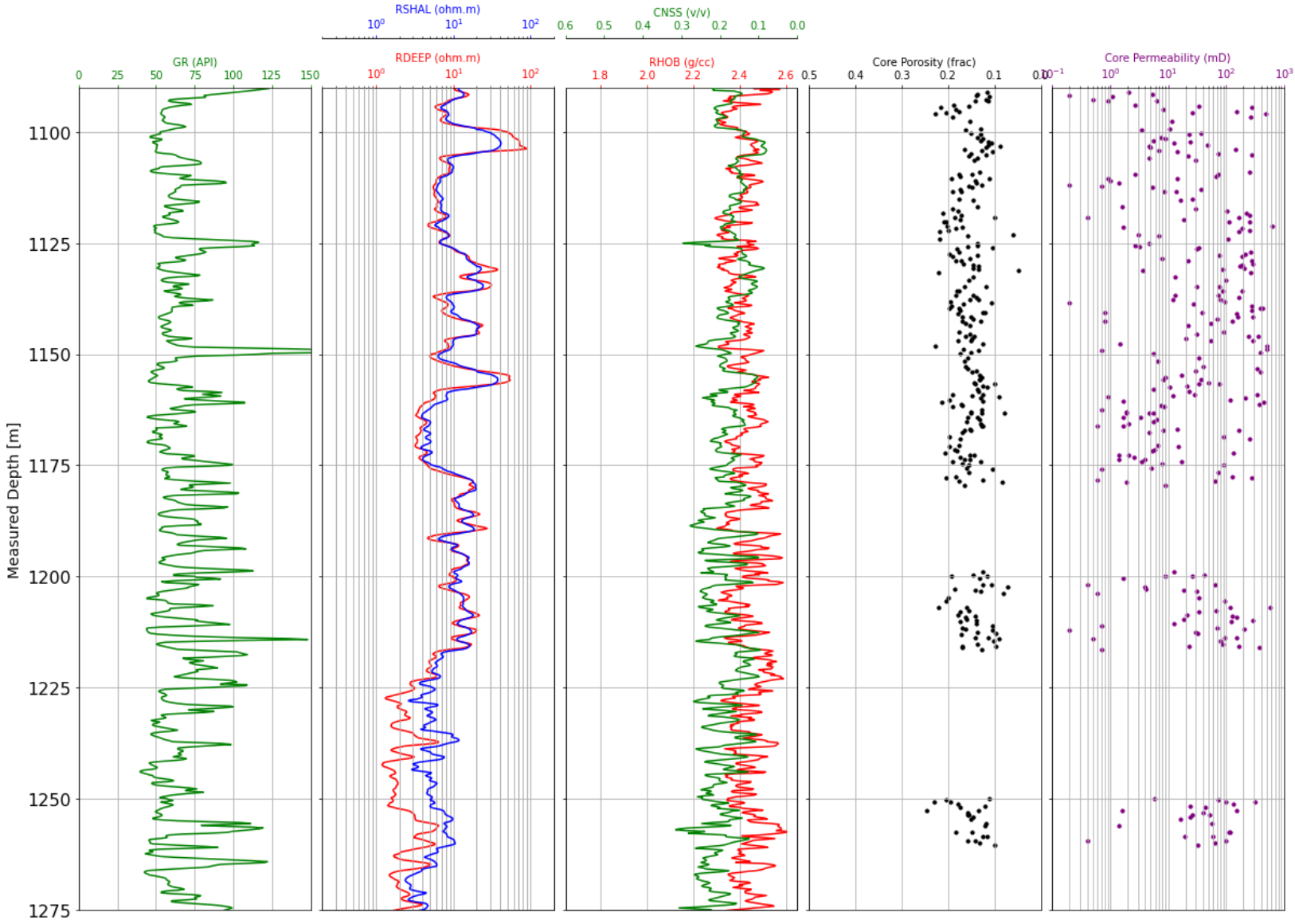
Nota. Elaboración propia

Por último, en la figura 26 se muestra el conjunto de registros y análisis de núcleos para el pozo ***7-RO-14-BA***. En la figura 26 se puede apreciar primeramente el cambiante comportamiento del GR a lo largo del registro, indicando la presencia de minerales radioactivos en pequeños intervalos, además, los cuerpos de baja radioactividad presentes no son aparentemente tan limpios como los observados en otros casos, teniendo un GR mínimo de 50 gAPI aproximadamente. Adicionalmente, los valores del registro neutrón no son significativamente altos (mayor a 0.4) en los intervalos con alto valor de GR, pudiéndose tratar de minerales siliciclásticos con alta respuesta en radioactividad como feldespatos potásicos, sin embargo, debe validarse con más información mineralógica y petrofísica del campo y formación en estudio. Por último, se observan pequeños *crossovers* entre el registro densidad y neutrón indicando la posible presencia de gas predominantemente entre 1128 m y 1136 m. Se observa además la desviación apreciada de la figura 26 en el gran rango de valores que tiene la permeabilidad al aire a lo largo del intervalo corazonado.

Adicionalmente, según la información disponible para la formación Sergi en la página web, se tienen únicamente 45 pozos de la cuenca con registros disponibles para esta formación. Sin embargo, el campo Remanso, al cual pertenece el pozo ***7-RO-14-BA***, a pesar de tener reportados 122 pozos, solo uno tiene registros disponibles, siendo este el pozo analizado, estando los 121 pozos restantes del campo sin ningún tipo de data disponible.

Figura 26

Registros y análisis de núcleos para el pozo 7-RO-14-BA



Nota. Elaboración propia usando en el lenguaje Python.

Los pozos con registros disponibles en la formación Sergi más cercanos son los pozos ***3-GTE-4DPA-BA*** (campo TIÊ), ***6-BRSA-1326D-BA*** (campo Massape), ***4-BRSA-1060D-BA*** (campo Taquipe), ***4-BRSA-1186D-BA*** (campo Taquipe), ***3-BRSA-1177D-BA*** (campo Taquipe), y ***3-BRSA-1217-BA*** (campo Taquipe). Todos estos pozos se encuentran a una distancia aproximada de 16 km del pozo en estudio. Consultando la columna estratigráfica disponible para

estos pozos cercanos, se encuentra primeramente que la formación Sergi en estos pozos se encuentra a una profundidad de por lo menos el doble de lo reportada para el pozo en estudio, y, en segundo lugar, se observa de igual manera la alta heterogeneidad vertical demarcada por el GR. Según lo referenciado en la sección 2.1.1, la cuenca Recóncavo cuenta con múltiples sistemas de trampas y sistemas petrolíferas que comparten las formaciones generadoras y almacén, variando la profundidad de estas en cada sistema. Esta condición hace difícil mantener la consistencia de los modelos obtenidos bajo la aplicación de la técnica sobre otros pozos disponibles de la formación Sergi, dado la variación que pueden tener los factores que afectan las propiedades que condicionan el flujo y almacenamiento en la roca, tales como mineralogía, textura, y factores geológicos. Por esta razón, se determina el caso de estudio de la formación Sergi en el presente trabajo como el desarrollo de la metodología para la data del pozo en estudio ***7-RO-14-BA***, del campo Remanso.

4. Modelo de Porosidad y Permeabilidad

Los modelos de porosidad y permeabilidad propuestos están compuestos por la porosidad total, la porosidad efectiva y la permeabilidad al aire. Según la metodología planteada por Amaefule, cada unidad de flujo hidráulica identificada tendrá su relación íntegra de porosidad efectiva y permeabilidad, por lo que la identificación de las unidades de flujo es el paso inicial para establecer dichos modelos. Para ello, se llevan a cabo análisis de *clusters* y usos de histogramas como plantea Amaefule, con el fin de realizar el mejor agrupamiento posible de la data de núcleo disponible. Consecutivamente, se obtienen los modelos de porosidad al aire y efectiva de los registros, los cuales deben correlacionarse con los datos de porosidad del núcleo y evaluar su error. En cuanto a la porosidad del núcleo, se tiene reportado que las pruebas en laboratorio fueron llevadas a cabo con el método del porosímetro de Helio. El gas inyectado ocupa sólo los poros conectados, y se considera una medición de la porosidad efectiva, sin embargo, el resultado dependerá del proceso de limpieza y secado de la muestra, por lo que en la práctica la medición varía de la porosidad total a porosidad efectiva, según la extracción del agua asociada a la arcilla. El modelo de permeabilidad se determina a partir de las correlaciones entre la permeabilidad al aire y porosidad efectiva del núcleo para cada unidad de flujo de manera independiente. Una vez la porosidad total o efectiva del registro esté validada con el núcleo, las unidades estarán caracterizadas con sus modelos respectivos junto con sus valores medios de porosidad efectiva, permeabilidad al aire e indicador de zona de flujo.

4.1 Caso Agua Grande-Jandaia

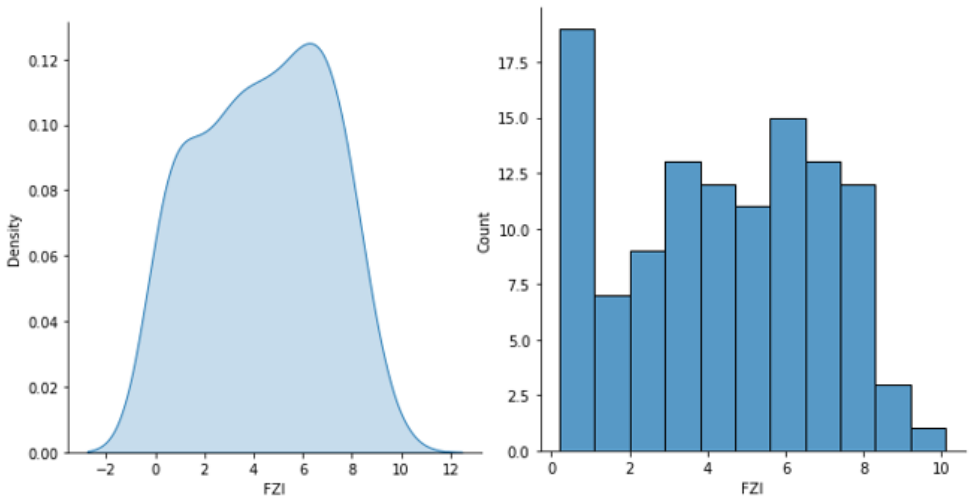
Para el caso de estudio determinado para la formación Agua Grande en el campo Jandaia, se toma inicialmente el conjunto de datos de análisis de núcleo de ambos pozos mostrado en la figura 25 y se determinan las unidades de flujo presentes.

Se escoge la librería Scikit-Learn de Python como la herramienta para llevar a cabo los análisis de *clusters* requeridos. Esta es una librería de aprendizaje automático de software libre que cuenta con varios algoritmos de clasificación, regresión y agrupamiento que está diseñada para interactuar con las librerías numéricas y científicas de Python NumPy y SciPy. Se escogen los modelos de mezcla gaussiana para hallar las unidades a partir de la distribución del indicador de zona de flujo calculado de los análisis de núcleos, teniendo 115 datos. Un modelo de mezcla gaussiana es un modelo probabilístico que supone que todos los puntos de datos se generan a partir de una mezcla de un número finito de distribuciones gaussianas con parámetros desconocidos. De esta forma, se podría considerar a los modelos mixtos como una generalización del agrupamiento mediante *K-Means*, para incorporar información sobre la estructura de covarianza de los datos, así como los centros de las gaussianas latentes. Particularmente, se usa el objeto *GaussianMixture* de la librería, para implementar el algoritmo de maximización de expectativas para ajustar modelos mixtos de Gauss. Se proporciona un método *GaussianMixture.fit* que aprende un modelo de mezcla gaussiana a partir de los datos del tren. Dados los datos de prueba, puede asignar a cada muestra el Gaussiano al que probablemente pertenece utilizando el método *GaussianMixture.predict.* Sin embargo, al igual que la mayoría de los métodos de agrupamiento no supervisado, el algoritmo requiere como parámetro inicial el número de *clusters* a agrupar. En las referencias de la librería Scikit-Learn existen métodos gráficos para seleccionar el número de

clusters óptimo, como por ejemplo el método del codo o *elbow method*, determinación de coeficientes de silueta, criterio de información de Akaike (AIC), criterio de información de Bayes (BIC). Sin embargo, la respuesta de estos algoritmos no es del todo concreta y siempre está condicionada por la calidad de los datos y su interpretación. En la figura 27 se muestra el comportamiento estadístico del FZI a agrupar, en donde se observa claramente la dispersión producto de múltiples modas y unidades de flujo identificadas en la figura 21. En la figura 28 se muestran los resultados del método del codo y el análisis de silueta aplicados al conjunto de datos de la figura 27, sugiriendo la elección de 3 grupos o *clusters*, tomando un valor óptimo de coeficiente de silueta cercano al codo del método gráfico. Finalmente, en la figura 29 se muestra el conjunto de datos del FZI clasificado.

Figura 27

Comportamiento estadístico FZI en Fm Agua Grande Campo Jandaia

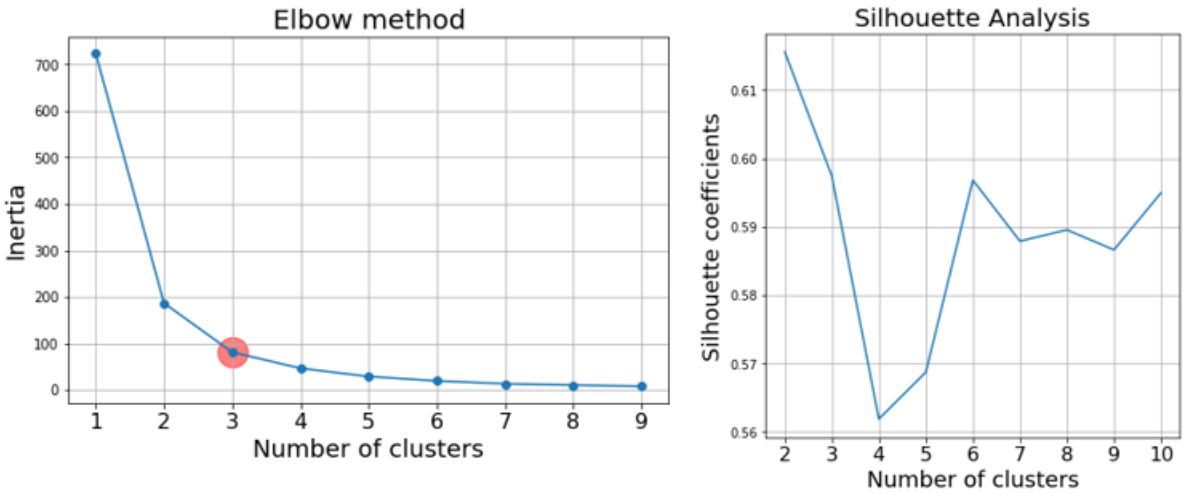


Nota. Elaboración propia.

En la figura 30 se muestra la relación logarítmica entre el RQI y la porosidad normalizada del campo Jandaia clasificando el conjunto de datos en 3 unidades.

Figura 28

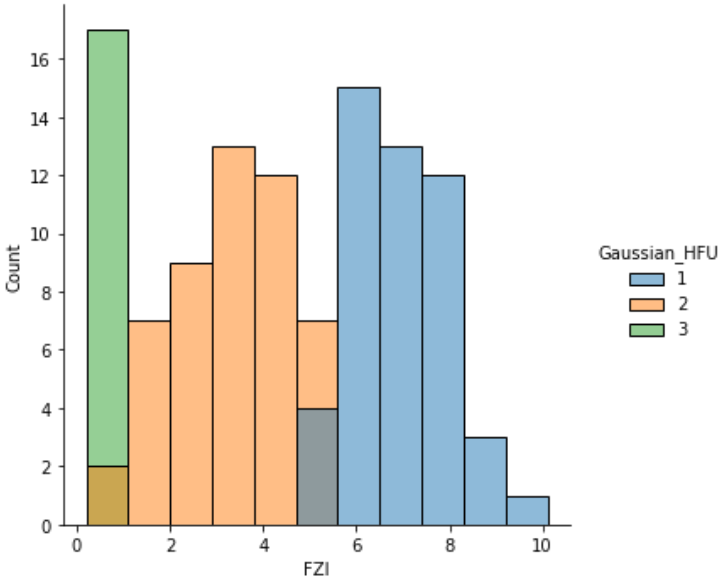
Método del codo y análisis de silueta sobre la data del FZI a agrupar



Nota. Elaboración propia.

Figura 29

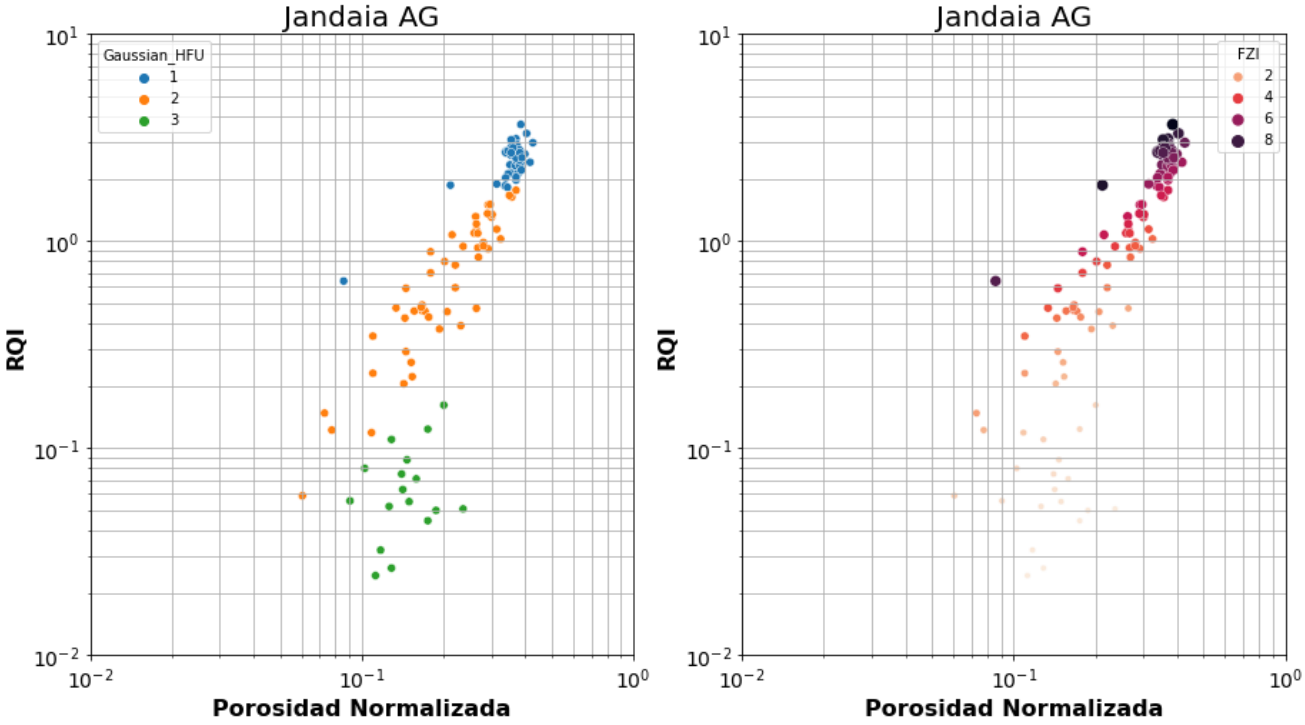
Comportamiento estadístico FZI clasificado en Fm Agua Grande Campo Jandaia



Nota. Elaboración propia.

Figura 30

Relación entre RQI y Porosidad normalizada para el campo Jandaia Fm Agua Grande

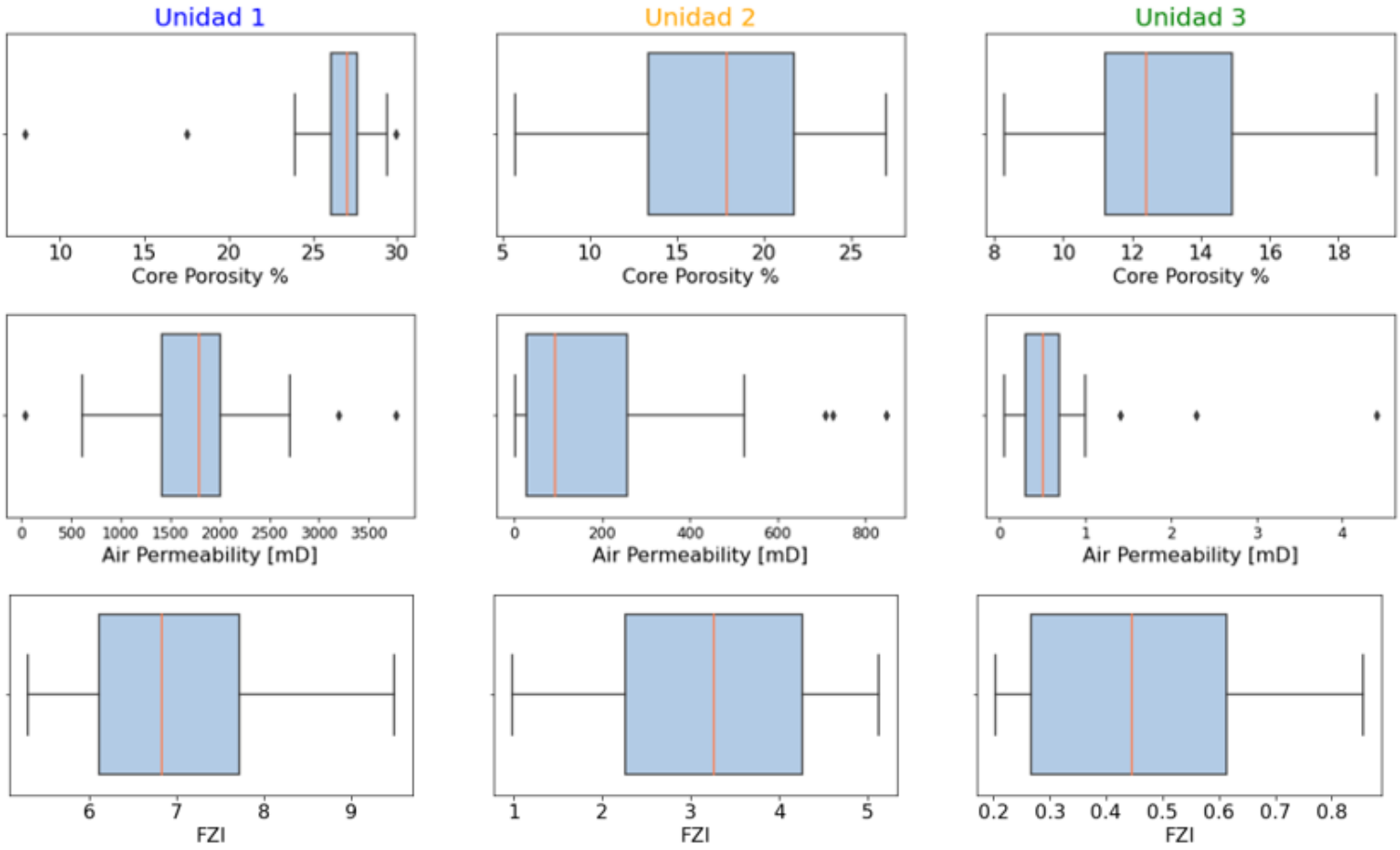


Nota. Elaboración propia.

La figura 31 muestra los diagramas de cajas y bigotes de la porosidad del núcleo, permeabilidad al aire y FZI en las 3 unidades identificadas. Primeramente, se aprecia que la unidad 1 representa la mejor unidad de flujo en base a sus propiedades, teniendo la mejor calidad de roca con una mediana de porosidad de 27% , permeabilidad al aire de 1794 mD y FZI de 6.82. Segundo, la unidad de flujo 2 se ubica como la unidad intermedia en cuanto a calidad de roca con una mediana de porosidad de 17.9%, permeabilidad al aire de 91 mD y FZI de 3.26. Por último, la unidad de flujo 3 representa la peor calidad de roca con una mediana de porosidad de 12.4%, permeabilidad al aire de 0.5 mD y FZI de 0.44.

Figura 31

Diagrama de cajas y bigotes para las unidades identificadas en la Fm Agua Grande



Nota. Elaboración propia.

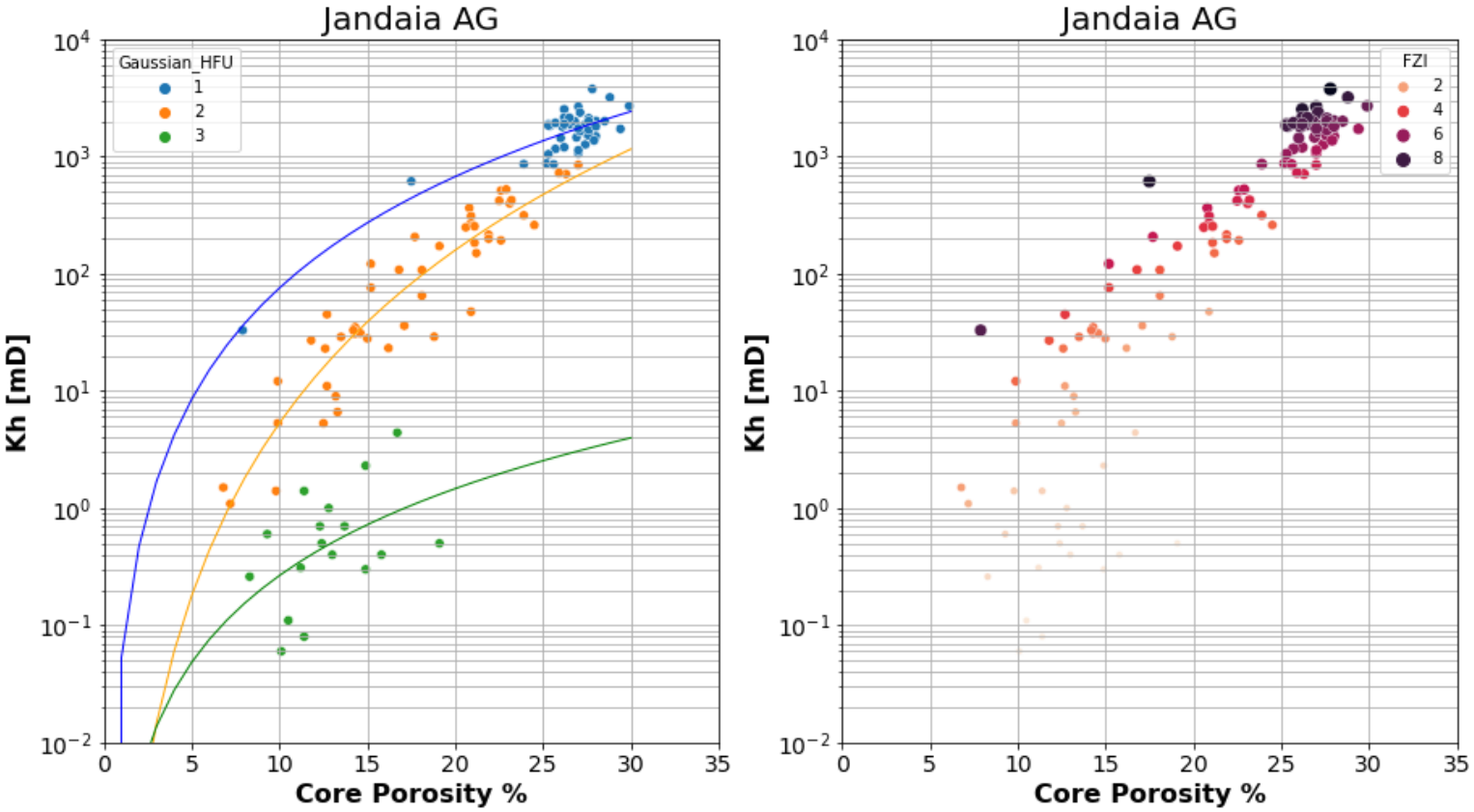
En la figrua 32 se observa el *crossplot* entre la permeabilidad al aire y la porosidad del núcleo con las correlaciones desarrolladas para las unidades de flujo identificadas.

El último paso para el presente caso de estudio consiste en determinar los modelos de porosidad total y efectiva de los intervalos analizados a partir de los registros disponibles. La determinación de estos modelos está influenciada por el tipo de arcilla presente, ya que se establece que la diferencia entre estos modelos consiste en la consideración del agua adherida a los minerales de arcilla. Por ende, la cuantificación de arcilla (Vcl) y la determinación de los *end points* (lectura

de densidad y neutrón asociada a la arcilla) es un factor primordial. Con el objetivo de facilitar el procedimiento, se usa el software especializado Interactive Petrophysics (IP) para el cálculo del Vcl y establecer los *end ponits* respectivos de manera gráfica. El Vcl es calculado como el valor mínimo entre el Vcl del GR y de la combinación densidad-neutrón. De esta forma, se pueden evitar errores al considerar zonas de alta radioactividad como arcillosas cuando está la posibilidad de que algún otro mineral esté haciendo tal efecto.

Figura 32

Crossplot de permeabilidad al aire y porosidad efectiva con los modelos de las unidades de flujo



Nota. Elaboración propia.

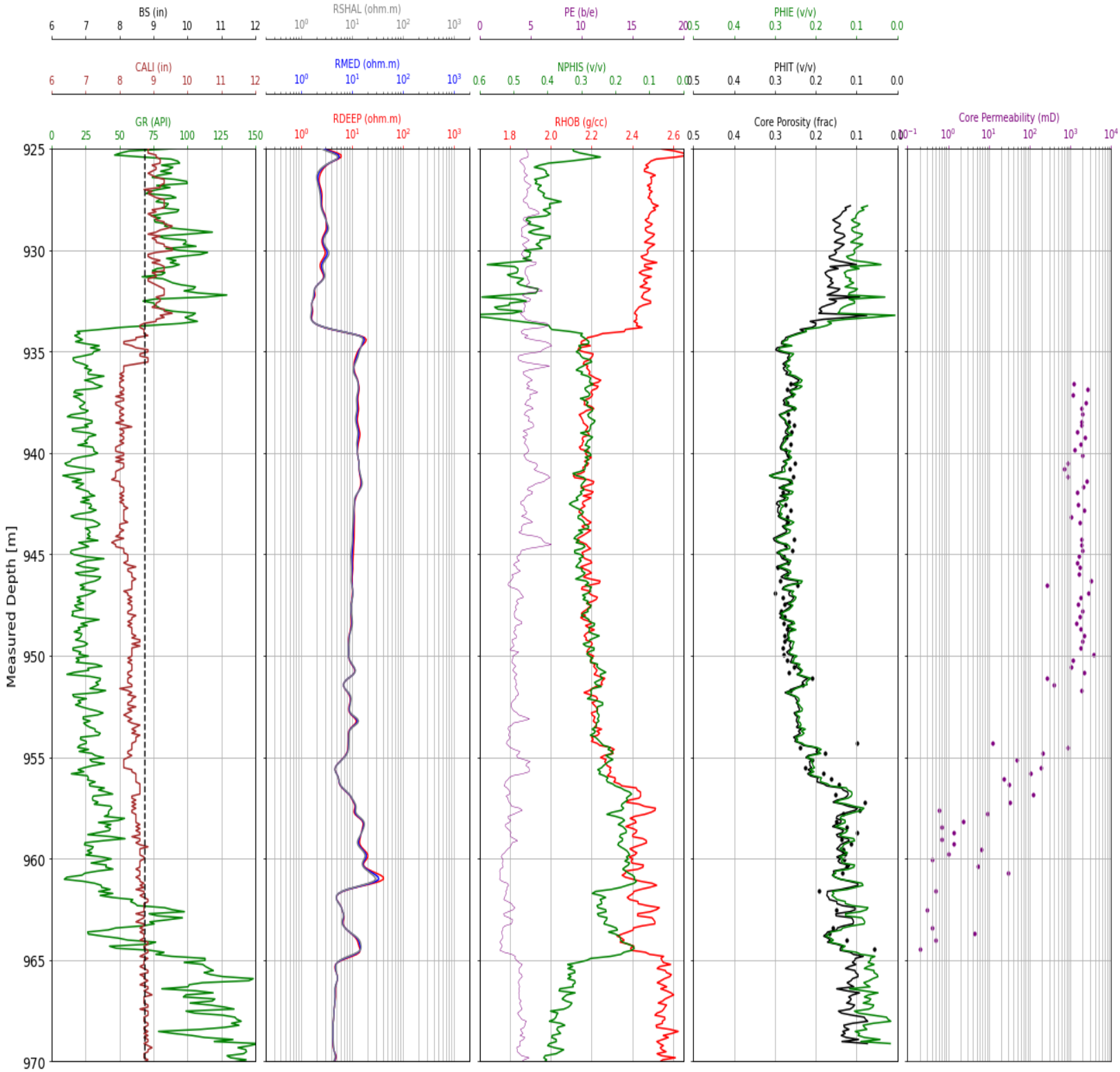
Para el primer pozo analizado, ***7-JND-3D-BA***, se establece un GR limpio de 7 gAPI y de arcilla de 173 gAPI. Así mismo, se fija el *wet clay point* como densidad de 2.658 g/cc y porosidad neutrón 0.354. Para el segundo pozo analizado, ***7-JND-13D-BA***, se establece un GR limpio de 7 gAPI y de arcilla de 121 gAPI. Así mismo, se fija el *wet clay point* como densidad de 2.624 g/cc y porosidad neutrón 0.21.

Teniendo esto en cuenta y al no tener información sobre la mineralogía del intervalo corazonado, la respuesta relativamente baja del neutrón al *wet clay* puede indicar que la arcilla presente en el intervalo no posee un alto valor de porosidad asociada a su agua adherida, y dado que el intervalo es considerablemente limpio como se ha mencionado anteriormente, la diferencia entre la porosidad total y efectiva calculadas no será alta.

Las figuras 33 y 34 muestran los modelos de porosdiad efectiva y total obtenidos para los dos pozos de Agua Grande. Se puede considerar una correlación aceptable entre la porosidad del núcleo y las calculadas. Además, como se comentó, se observa una muy pequeña diferencia entre la porosidad efectiva y total de los registros. La porosidad total del registro tiene un mejor ajuste lineal a la porosidad de Helio, sin embargo, dado que la diferencia con la porosidad efectiva calculada no es significativa y no se tienen los parámetros de laboratorio con los que fue medida la porosidad, el valor reportado del núcleo será considerado como porosidad efectiva.

Figura 33

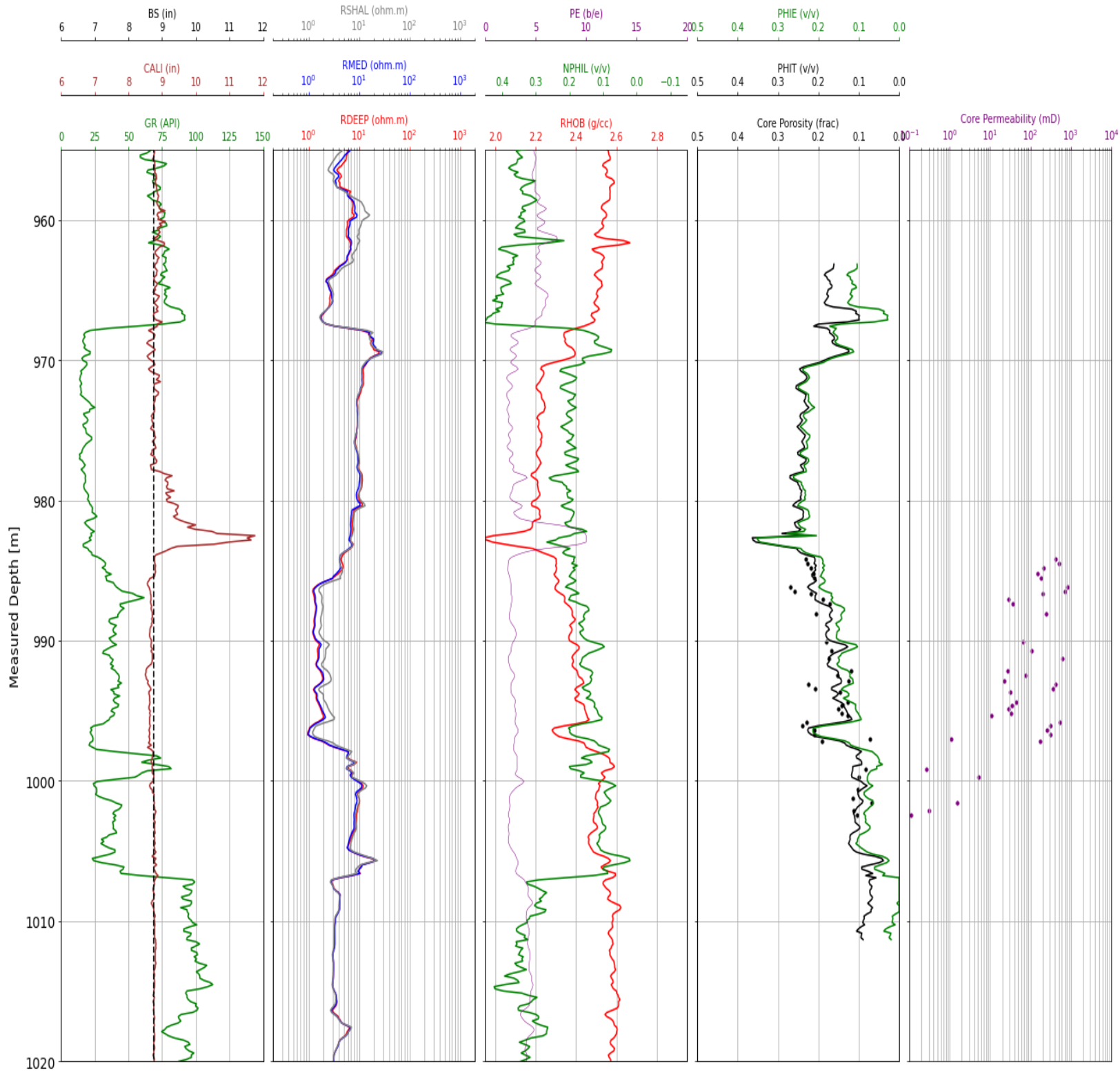
Registros del pozo 7-JND-3D-BA con los modelos de porosidad total y efectiva



Nota. Elaboración propia

Figura 34

Registros del pozo 7-JND-13D-BA con los modelos de porosidad total y efectiva



Nota. Elaboración propia

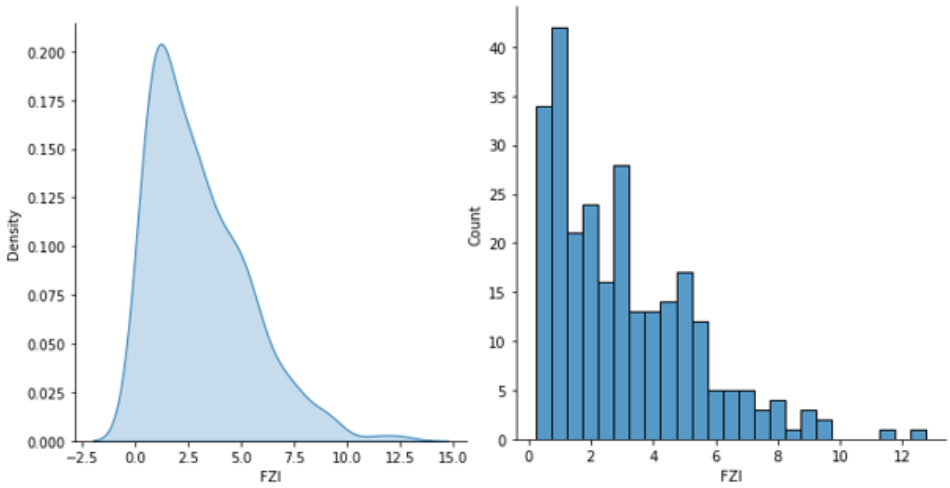
4.2 Caso Sergi-Remanso

Para el caso de estudio determinado para la formación Sergi del campo Remanso, se toma inicialmente el conjunto de datos de la figura 25 y se determinan las unidades de flujo presentes. El algoritmo propuesto para realizar el *clustering* sobre el indicador de zona de flujo del pozo ***7-RO-14-BA*** es el mismo que para el caso Agua Grande-Jandaia.

El caso Sergi-Remanso es considerado en la presente investigación como el más desafiante y donde mayor incertidumbre se tiene al realizar la metodología propuesta. La alta heterogeneidad vertical encontrada en el análisis exploratorio de datos junto con la alta dispersión de los datos de permeabilidad al aire lo hacen complejo a la aplicación de los algoritmos propuestos principalmente debido a la data disponible para modelar las unidades presentes, y esto hace que aspectos mineralógicos y texturales (tipo y distribución de arcilla en el medio poroso, minerales siliciclásticos presentes) que posiblemente influyan significativamente en el valor de la permeabilidad absoluta, no puedan ser tenidos en cuenta en los modelos predictivos del siguiente capítulo. En la figura 35 se observa el comportamiento estadístico del FZI a agrupar. Al aplicar el método del codo y el análisis de silueta en el conjunto de datos de la figura 35 se obtienen 4 grupos o *clusters* como el número óptimo a utilizar (figura 36).

Figura 35

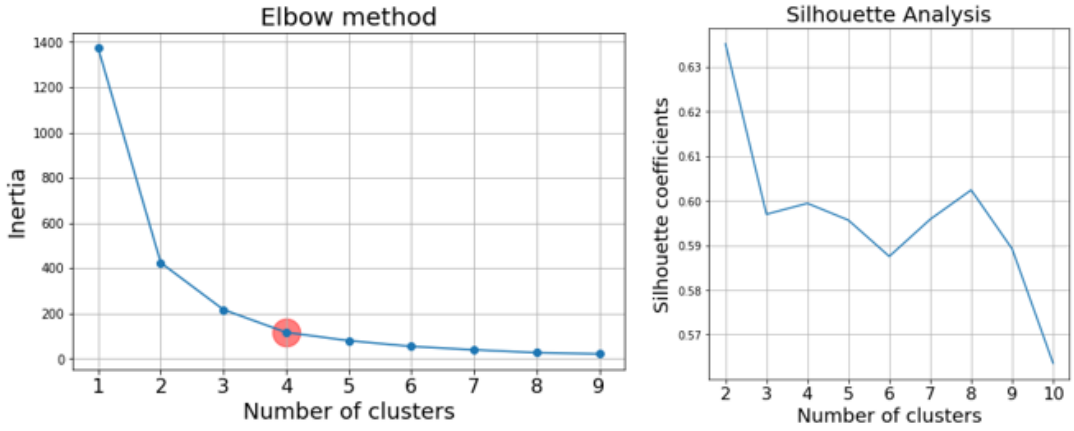
Comportamiento estadístico FZI en Fm Sergi Campo Remanso



Nota. Elaboración propia

Figura 36

Método del codo y análisis de silueta sobre la data del FZI a agrupar

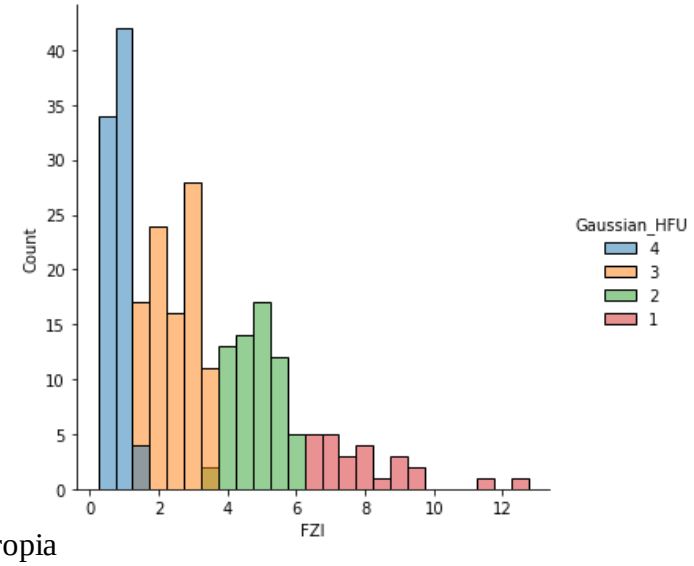


Nota. Elaboración propia.

Aplicando el algoritmo de agrupación del caso anterior (*Gaussian Mixture*) con 4 *clusters*, se obtienen los resultados de la figura 37. En la figura 38 se muestra la relación entre el RQI y la porosidad normalizada con las unidades identificadas.

Figura 37

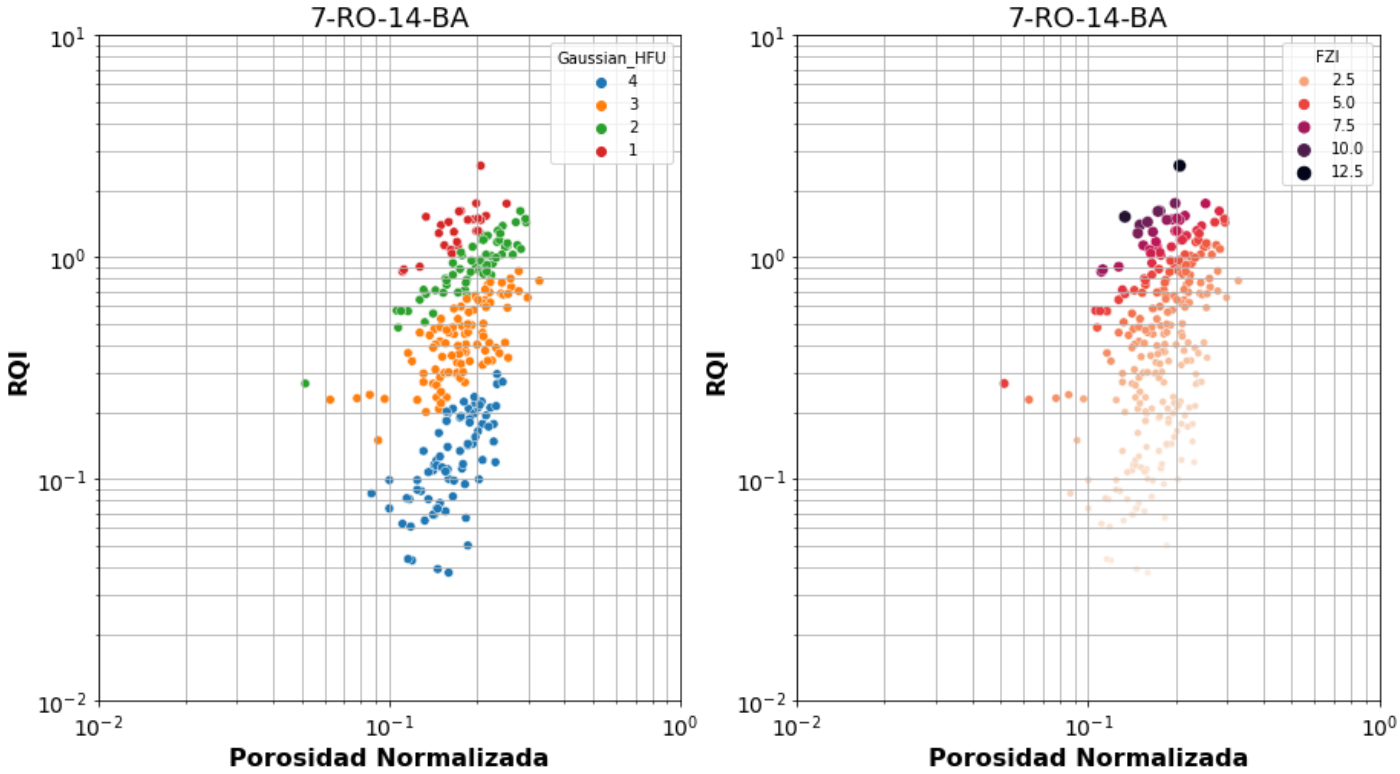
Comportamiento estadístico FZI clasificado en Fm Sergi campo Remanso



Nota. Elaboración propia

Figura 38

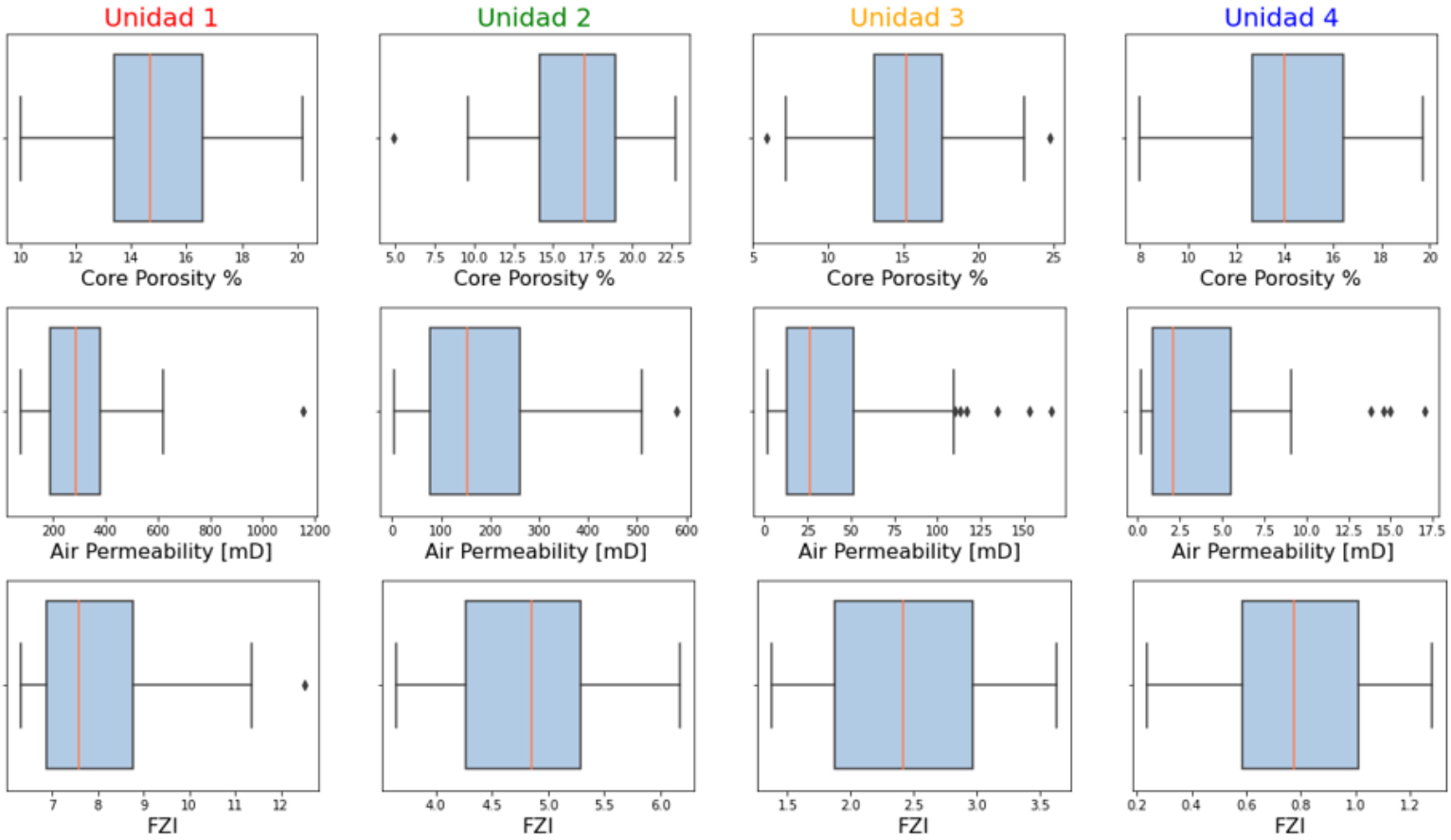
Relación entre RQI y Porosidad normalizada para el caso Sergi-Remanso



Nota. Elaboración propia

Figura 39

Diagrama de cajas y bigotes para las unidades identificadas en la Fm Sergi



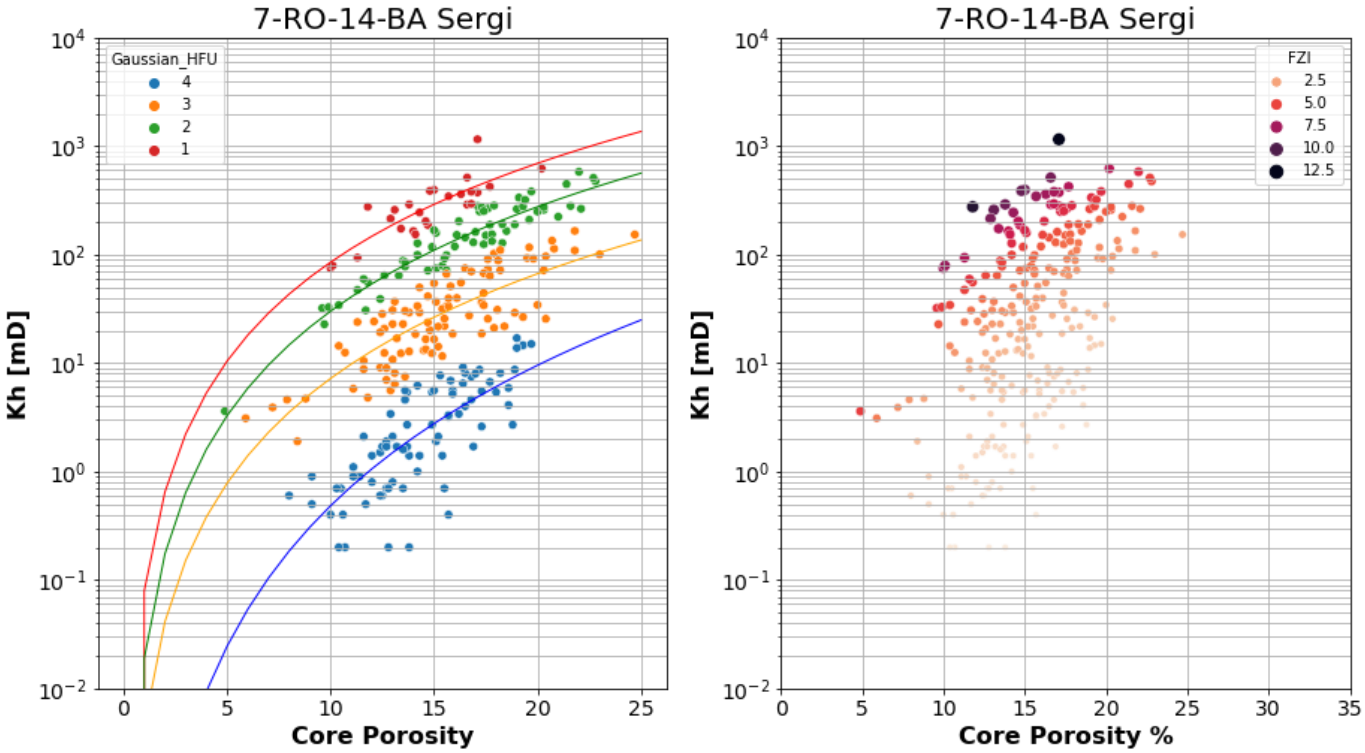
Nota. Elaboración propia

La figura 39 muestra los diagramas de cajas y bigotes de la porosidad del núcleo, permeabilidad al aire y FZI en las 4 unidades identificadas. Primeramente, se aprecia que la unidad 1 representa la mejor unidad de flujo en base a sus propiedades, teniendo la mejor calidad de roca con una mediana de porosidad de 14.7% , permeabilidad al aire de 289 mD y FZI de 7.59. Segundo, la unidad de flujo 2 se ubica como la unidad intermedia en cuanto a calidad de roca con una mediana de porosidad de 17%, permeabilidad al aire de 151 mD y FZI de 4.85. Tercero, la unidad de flujo 3 representa la unidad regular en cuanto a calidad de roca con una mediana de porosidad de 15.1%, permeabilidad al aire de 26 mD y FZI de 2.42. Por último, la unidad de flujo 4 representa

la peor calidad de roca con una mediana de porosidad de 14%, permeabilidad al aire de 2.1 mD y FZI de 0.77. En la figrua 40 se observa el *crossplot* entre la permeabilidad al aire y la porosidad del núcleo con las correlaciones desarrolladas para las unidades de flujo identificadas.

Figura 40

Crossplot de permeabilidad al aire y porosidad con los modelos de las unidades de flujo

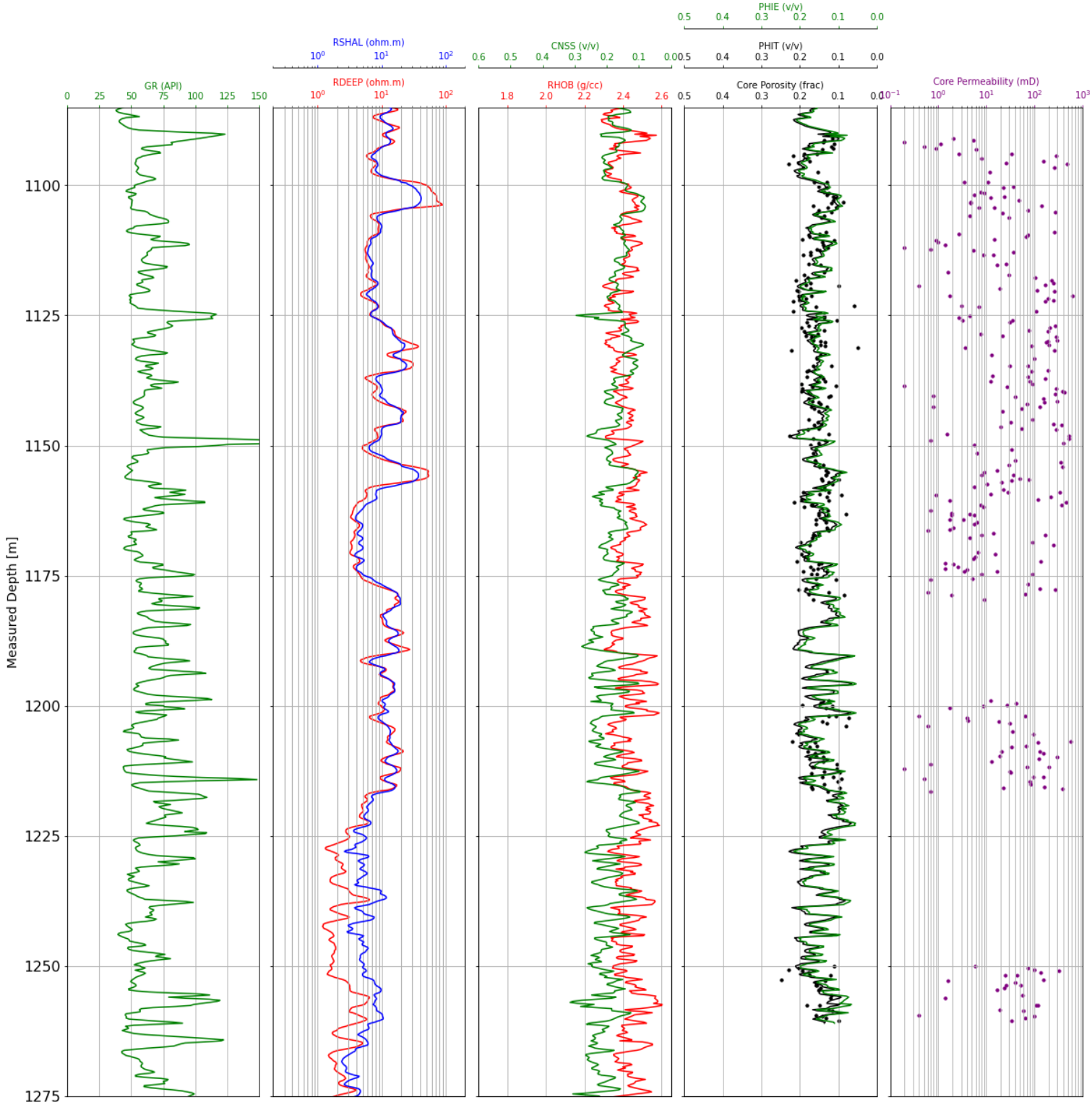


Nota. Elaboración propia

La determinación de los modelos de porosidad total y efectiva del pozo, al igual que en el caso anterior, está dominado por los aspectos mineralógicos presentes en el intervalo analizado. Sin embargo, dada la alta heterogeneidad vertical del intervalo y la poca data disponible asociada a la mineralogía, los *end points* fueron ajustados para obtener el mejor *match* con los datos de núcleos, siendo el valor de densidad *wet clay* 2.750 g/cc y porosidad neutrón de 0.3 v/v. La figura 41 muestra los modelos obtenidos. Al igual que el caso anterior, la porosidad del núcleo se considerará como efectiva.

Figura 41

Registros del pozo 7-RO-14-BA con los modelos de porosidad total y efectiva



Nota. Elaboración propia

5. Predicción de unidades de flujo usando *Machine Learning*

En el presente capítulo se establecen dos paradigmas del *Machine Learning* en la determinación de unidades de flujo para ambos casos de estudio, el aprendizaje supervisado y no supervisado. En cuanto al primero, se propone la aplicación de algoritmos de regresión, en los cuales se busca modelar el indicador de zona de flujo en base al comportamiento de los registros de pozo, y algoritmos de clasificación, en donde se busca identificar directamente la unidad de flujo a partir de los registros. Para el aprendizaje no supervisado, se realizan técnicas de *clustering* o agrupamiento de los registros en búsqueda de la identificación de las unidades de flujo hidráulicas verdaderas del capítulo anterior.

Uno de los factores claves al momento de desarrollar algoritmos de *Machine Learning* consiste en estructurar adecuadamente el flujo de trabajo para evitar el sobreajuste y subajuste de los modelos supervisados (estimadores), así como optimizar los parámetros de entrada de cada uno, conocidos como hiperparámetros. El flujo de trabajo llevado a cabo en la presente investigación se muestra en la figura 42, sin embargo, la separación de la data en el set de evaluación y validación dependerá según el modelo.

En términos generales, se divide inicialmente el set de datos en un set de prueba y un set de entrenamiento. Sin esto, los modelos podrían llegar a obtener un puntaje perfecto, pero fallarían al momento de predecir resultados útiles sobre datos nuevos (sobreajuste). La función *train_test_split* de Scikit-Learn es usada para este propósito. Sin embargo, al evaluar diferentes configuraciones (hiperparámetros) para los estimadores, aún existe el riesgo de sobreajuste en el conjunto de prueba porque los parámetros se pueden modificar hasta que el estimador funcione de

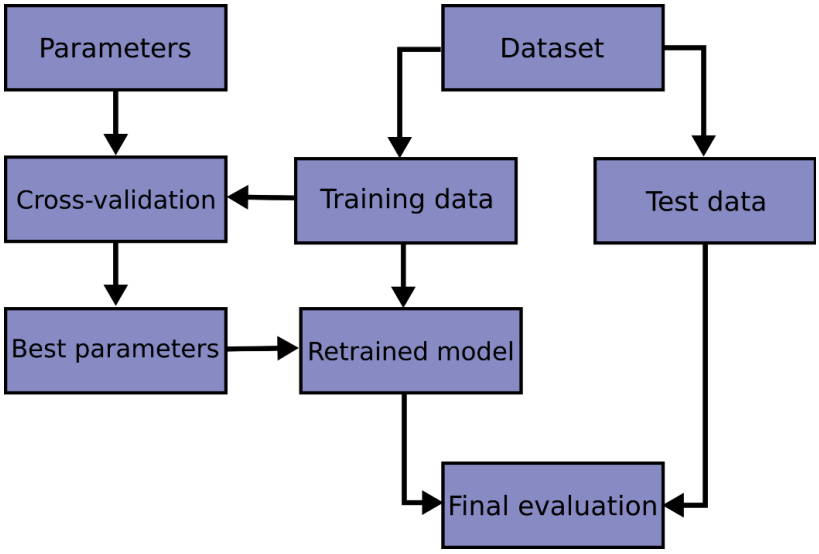
manera óptima. De esta manera, el conocimiento sobre el conjunto de prueba puede filtrarse en el modelo y las métricas de evaluación ya no informan sobre el rendimiento de la generalización.

Para resolver este problema, se puede presentar otra parte del conjunto de datos como el llamado conjunto de validación: el entrenamiento continúa en el conjunto de entrenamiento, después de lo cual se realiza la evaluación en el conjunto de validación y cuando el experimento parece tener éxito, la evaluación final se puede hacer en el conjunto de prueba. Sin embargo, al dividir los datos disponibles en tres conjuntos, se reduce drásticamente la cantidad de muestras que se pueden usar para aprender el modelo, y los resultados pueden depender de una elección aleatoria particular para el par de conjuntos (entrenamiento, validación). Una solución a este problema es un procedimiento llamado validación cruzada (CV para abreviar). Todavía se debe reservar un conjunto de prueba para la evaluación final, pero el conjunto de validación ya no es necesario al hacer CV. En el enfoque básico, llamado *k -fold* CV, el conjunto de entrenamiento se divide en k conjuntos más pequeños (figura 43).

La medida de rendimiento informada por la validación cruzada de k veces es entonces el promedio de los valores calculados en el ciclo. Este enfoque puede ser computacionalmente costoso, pero no desperdicia demasiados datos (como es el caso cuando se arregla un conjunto de validación arbitrario), lo cual es una gran ventaja en problemas como la inferencia inversa donde el número de muestras es muy pequeño.

Figura 42

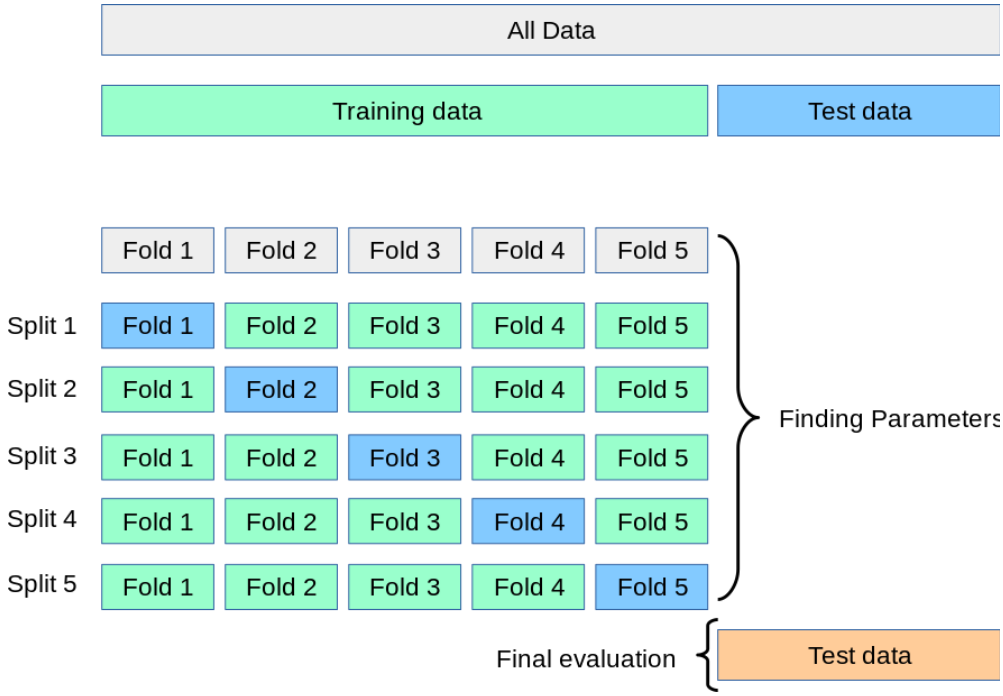
Flujo de trabajo en la aplicación de Machine Learning



Nota. Tomado de https://scikit-learn.org/

Figura 43

Técnica de validación cruzada



Nota. Tomado de https://scikit-learn.org/

El primer estimador de aprendizaje supervisado a implementar será una regresión lineal múltiple (siendo las variables independientes los registros disponibles), minimizando la suma residual de cuadrados entre las observaciones en el conjunto de datos (FZI) y los objetivos predichos. Otras regresiones lineales como la regresión de cresta o la regresión de Lasso no serán tenidas en cuenta dado que cuentan con grandes similitudes de cálculo con la regresión lineal clásica. El siguiente estimador en esta categoría son las máquinas de vectores de soporte (SVM) en sus categorías de regresión y clasificación. Una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a dos espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase. Finalmente, se usarán parte de los estimadores más poderosos dentro del *Machine Learning* que están agrupados en una categoría llamada métodos de aprendizaje en conjunto: *Random Forest*, y *GradientBoost*. El primero consiste en un método de promediación, cuyo principio fundamental es construir varios estimadores de forma independiente y luego promediar sus predicciones. En general, el estimador combinado suele ser mejor que cualquiera de los estimadores de base única porque se reduce su varianza. Por otro lado, el estimador *GradientBoost* es un método de *Boosting* o impulsos, en donde los estimadores bases se construyen secuencialmente y se intenta reducir el sesgo del estimador combinado. La motivación es combinar varios modelos débiles para producir un conjunto poderoso.

En cuanto al aprendizaje no supervisado, se usará primeramente el estimador *K-Means*, el cual agrupa los datos tratando de separar muestras en n grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo. Este algoritmo requiere que

se especifique el número de *clusters*. El siguiente estimador a utilizar será el *Hierarchical Clustering*, el cual es una familia de algoritmos de agrupamiento que crean *clusters* anidados fusionándolos o dividiéndolos sucesivamente. Esta jerarquía de conglomerados se representa como un árbol (o dendrograma). La raíz del árbol es el único racimo que reúne todas las muestras, siendo las hojas los racimos con una sola muestra.

A continuación, se describen los detalles de la aplicación de cada estimador para los dos casos de estudio planteados en la investigación y sus principales consideraciones. El análisis y escalamiento de los modelos desarrollados usando los registros disponibles se lleva a cabo en el siguiente capítulo.

5.1 Caso Agua Grande-Jandaia

Para este primer caso de estudio se declaran inicialmente las variables de entrada como la data de los registros disponibles puestos a profundidad con los cálculos del indicador de zona de flujo y la clasificación de las unidades de flujo. Los registros usados fueron el GR, Densidad, y Neutrón de las figuras 15 y 20. En primera instancia, se combina la data de ambos pozos para construir el *dataset* más robusto posible y compensar la poca cantidad de datos disponibles. Segundo, como se observa en la figura 20, los datos de núcleos del segundo pozo no cubren la totalidad del intervalo, por lo que si se usan de forma individual no reflejará el comportamiento íntegro de la formación.

El primer estimador a usar es la regresión lineal múltiple. Para este caso, no se requiere de estandarización en los datos ni fragmentación del *dataset* en set de prueba y entrenamiento. Al modelar el indicador de zona de flujo con los registros mencionados, se obtiene un coeficiente de

determinación de 0.62 y un error cuadrado medio de 2.455 (sensible a la magnitud del indicador de zona de flujo). Según las métricas de medición del modelo, existe una desviación importante entre el FZI original y el predicho, por lo que en la identificación de las unidades de flujo bajo este enfoque no se obtendrá una precisión aceptable.

El siguiente estimador aplicado es el regresor de vectores de soporte (SVR). Para este algoritmo y los siguientes, se usa el método de estandarización para escalar los datos. Si una característica tiene una varianza de órdenes de magnitud mayor que otras, podría dominar la función objetivo y hacer que el estimador no pueda aprender de otras características correctamente como se esperaba. Por lo tanto, al aplicar la estandarización los datos de entrada son escalados para tener media cero y varianza unitaria. En este caso, se procede a utilizar el set de entrenamiento para realizar el aprendizaje, con un 70% de los datos, y el restante como set de prueba en el cual se obtienen las métricas del estimador. Adicionalmente, se prueba el estimador en el método de validación cruzada sobre 10 pliegues y se obtiene un coeficiente de determinación promedio de 0.48 y una desviación estándar del 29%. Esto indica que el modelo no está haciendo sobreajuste a los datos de entrenamiento, aunque obtiene pobres resultados. Al modelar el indicador de zona de flujo se obtiene un coeficiente de determinación de 0.59 y un error cuadrado medio de 2.69. Por lo tanto, el estimador no obtiene la precisión adecuada en la determinación del FZI.

El siguiente estimador a utilizar es el *Random Forest Regressor* (RFR), se usa el método de estandarización para escalar los datos. Al comprobar la técnica de validación cruzada con 5 pliegues sobre los datos de entrenamiento se obtiene un coeficiente de determinación promedio de 0.43 y una desviación estándar de 19%, por lo que el estimador a pesar de no entregar resultados aceptables no realiza sobreajuste a los datos, lo cual es muy importante para este tipo de estimador. Finalmente, el estimador sobre los datos de prueba obtiene un coeficiente de determinación de

0.617 y un error cuadrado medio de 2.511. En la tabla 10 se tabulan los resultados del coeficiente de determinación (r^2) y el error cuadrado medio (MSE) para los estimadores de regresión usados, teniendo en cuenta el modelamiento para el *set* de entrenamiento, de prueba, y toda la data. En el caso del *Random Forest Regressor*, solo se reportan los resultados del *set* de prueba, ya que en este tipo de estimadores si se evalúan los resultados sobre los datos que se usaron en el entrenamiento se obtendrá un sobreajuste, de manera que se consideran los resultados del *set* de prueba como la mejor aproximación a una nueva predicción.

Tabla 10

Resultados modelamiento de FZI con estimadores de regresión caso Agua Grande

	Training Dataset	*Test Dataset*	*All Data*
Linear Regression r^2			0.620
Linear Regression MSE			2.455
SVR r^2	0.618	0.59	0.610
SVR MSE	2.445	2.69	2.519
RFR r^2		0.617	
RFR MSE		2.511	

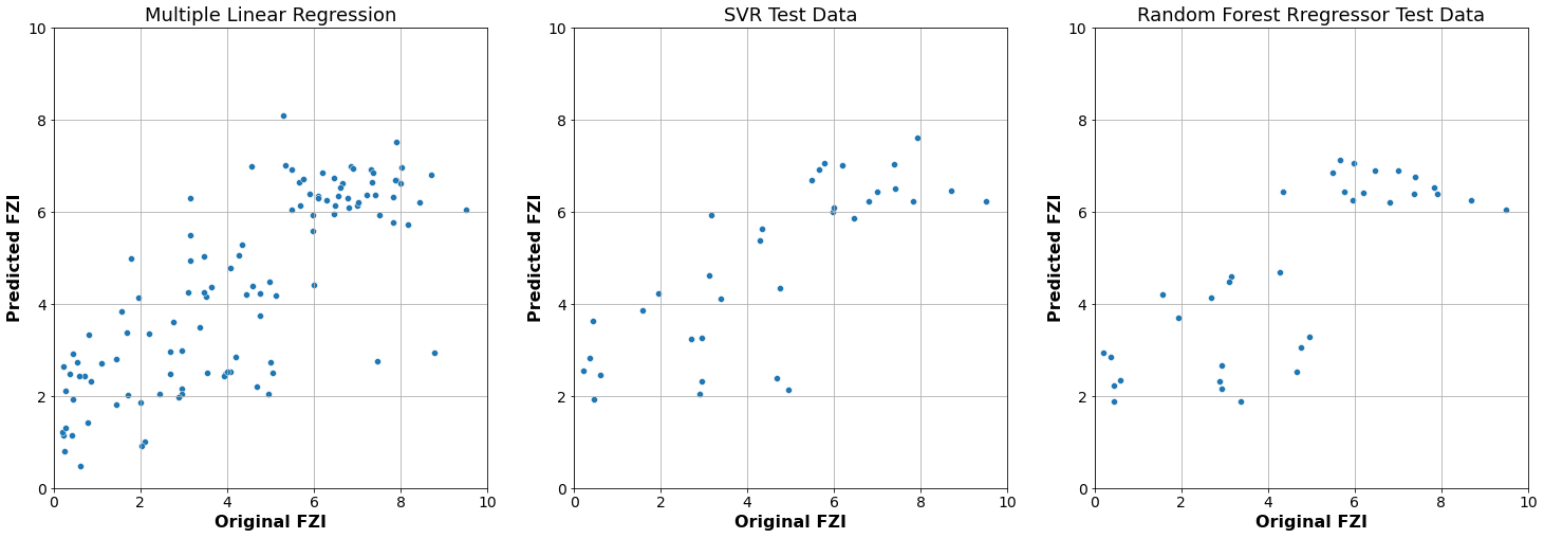
Nota. Elaboración propia

Según lo obtenido por los tres estimadores distintos usados en la regresión, se puede confirmar que este tipo de aprendizaje no es adecuado para modelar e identificar las unidades de flujo a partir del indicador de zona de flujo, tal como se preveía de los estudios previos comentados en los antecedentes investigativos. Esto principalmente se debe a alguna de las suposiciones de las regresiones usadas: linealidad, homocedasticidad, y falta de multicolinealidad principalmente, en las variables de entrada o también llamados regresores. Sin embargo, el algoritmo *Random Forest Regressor* representa la mejor aproximación para este caso, dado que, a pesar de tener un

coeficiente de determinación menor a la regresión lineal, la evaluación del modelo se realiza sobre datos nuevos no usados para el entrenamiento. En la figura 44 se resumen los resultados de los algoritmos de regresión para este caso de estudio.

Figura 44

Resultados regresiones caso Agua Grande



Nota. Elaboración propia

El siguiente tipo de estimadores corresponde a algoritmos de clasificación. Para estos, las variables de entrada son los mismos registros que el caso anterior pero las etiquetas serán las unidades de flujo identificadas en el capítulo anterior. En este caso, los modelos se evaluarán con el test de evaluación, teniendo este un 30% de manera que los modelos sean entrenados con el 70% restante. Así mismo, se usará el reporte de clasificación suministrado por la librería Scikit-Learn en donde se entregan las métricas más comunes en la evaluación del desempeño de los modelos de clasificación: exactitud (predicciones correctas sobre total de predicciones), precisión (verdaderos positivos sobre la suma de verdaderos positivos y falsos positivos), *recall* (verdaderos positivos sobre la suma de verdaderos positivos y falsos negativos), y *F1 score* (promedio entre precisión y *recall).*

El primer estimador a usar es el *Support Vector Classifier* (SVM) y el método de estandarización para escalar los datos. Se crea una malla de hiperparámetros para realizar la búsqueda mediante validación cruzada la mejor combinación de estos, teniendo en cuenta pesos balanceados para las clases presentes dado que la unidad de flujo 3 tiene menos datos para entrenar (los pesos balanceados por clase son calculados como el número de muestras sobre el número de clases por la frecuencia respectiva de casa clase). El resultado arroja un parámetro C de regularización de 10 y un kernel lineal (forma de vectores de soporte) con una exactitud o *accuracy* de 75%. Al probar el set de evaluación escalado en el modelo, se obtienen los resultados de la tabla 11. En esta, se destacan las 3 clases o unidades de flujo clasificadas, con sus respectivas métricas y frecuencia en el set de evaluación (*support*). Se destaca la unidad 1 como una clasificación correcta, en donde todas sus instancias fueron correctamente clasificadas y dentro de su clase no fue incluida ninguna otra. Panorama totalmente distinto a la clase 3, en donde si bien cuenta con un *recall* aceptable, la poca precisión indica que las instancias de esta clase fueron en su mayoría identificadas correctamente, pero otras clases fueron agregadas a esta erróneamente, siendo esta la unidad 2. Mediante la ejecución de una matriz de confusión multiclase, se observa que la poca precisión en la clasificación de la unidad 3 se debe a que 8 instancias fueron identificadas como unidad 3 por el modelo cuando en realidad no lo eran. Mientras que solo 1 instancia fue clasificada como unidad 2 por el modelo cuando en realidad no lo era, por ello su mayor precisión. Por lo tanto, se puede establecer que en futuras predicciones el modelo tenderá a clasificar erróneamente instancias de clase 2 como clase 3. Sin embargo, empieza a notarse la mejoría en resultados con respecto a los modelos de regresión.

Tabla 11

Reporte de Clasificación SVM – Agua Grande

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		1	1	1	13
2		0.88	0.47	0.61	15
3		0.38	0.83	0.53	6
	Accuracy			0.74	34
	Micro avg	0.75	0.77	0.71	34
	Weighted avg	0.84	0.74	0.74	34

Nota. Elaboración propia

El siguiente estimador a usar es el *Random Forest Classifier* y el método de estandarización para escalar los datos. La búsqueda de hiperparámetros con el set de entrenamiento y el método de validación cruzada, teniendo en cuenta el peso balanceado de las clases, entrega un número de estimadores óptimo de 6 (árboles de decisión), 2 características máximas a considerar para la división de los nodos de los árboles y 5 muestras mínimas requeridas para estar en un nodo de hoja, con un 78% de exactitud. Al probar el set de evaluación en el modelo, se obtienen los resultados de la tabla 12. En términos generales, los resultados son muy parecidos a los reportados para el modelo SVM. La principal diferencia radica en la mejora de la identificación en la clase o unidad 3, sin embargo, una instancia es incorrectamente predicha como clase 1, disminuyendo su *recall* en comparación con el modelo anterior.

Tabla 12

Reporte de Clasificación Random Forest Classifier -Agua Grande

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		1	0.92	0.96	13
2		0.80	0.53	0.64	15
3		0.42	0.83	0.56	6
	Accuracy			0.74	34
	Micro avg	0.74	0.76	0.72	34
	Weighted avg	0.81	0.74	0.75	34

Nota. Elaboración propia

Por último, se usa el estimador *Gradient Boost Classifier* y el método de estandarización para escalar los datos. Los pesos de las clases presentes son calculados intrínsecamente por el estimador. Se reportan hiperparámetros óptimos de tasa de aprendizaje de 0.3 y 10 estimadores (número de etapas de refuerzo a realizar), con una exactitud de 70%. Al probar el set de evaluación en el modelo, se obtienen los resultados de la tabla 13. De manera inicial, se aprecian las mejores métricas de evaluación en este modelo. Sin embargo, la clase 3 mantiene una métrica *F1-score* similar a la obtenida en el modelo anterior. Tal como se ha comentado a lo largo de la investigación, el número limitado de muestras influye sobre el aprendizaje de los modelos, por lo que con mayores muestras de la clase 3 se podrían obtener mejores métricas en su identificación. Sin embargo, se puede concluir que se observan resultados consistentes a pesar de esta limitante en los 3 modelos de clasificación evaluados.

Tabla 13

Reporte de Clasificación Gradient Boost Classifier – Agua Grande

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		1	0.92	0.96	13
2		0.76	0.87	0.81	15
3		0.60	0.50	0.55	6
	Accuracy			0.82	34
	Micro avg	0.79	0.76	0.77	34
	Weighted avg	0.83	0.82	0.82	34

Nota. Elaboración propia

Para el aprendizaje no supervisado, se declaran las variables de entrada como el set de registros estandarizado (GR, NPHI, RHOB) puestos a profundidad con las unidades identificadas tal como para el aprendizaje supervisado. El primer algoritmo implementado es *K-Means*¸ sin embargo, dado que de manera previa se conocen los *clusters* por las unidades identificadas, se usan 3 *clusters* en el algoritmo. Los resultados se muestran en la tabla 14.

Tabla 14

Reporte de Clasificación K-Means – Agua Grande

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		0.71	0.96	0.81	46
2		0.63	0.53	0.58	49
3		0.44	0.24	0.31	17
	Accuracy			0.66	112
	Micro avg	0.60	0.57	0.57	112
	Weighted avg	0.64	0.66	0.63	112

Nota. Elaboración propia

Como se puede observar de la tabla 14, con una exactitud o *accuracy* del 66%, el algoritmo tiene resultados menos favorables que cualquier otro de clasificación en aprendizaje supervisado, tal como lo revisado en los antecedentes investigativos. Esto se debe principalmente a dos motivos. El primero consiste en las variables de entrada, ya que al ser estos registros estandarizados con correcciones ambientales básicas, el algoritmo realiza el *clustering* bajo las respuestas de estos, asociando su clasificación a electrofacies, no a unidades de flujo específicamente, y tal como está expuesto por estudios anteriores en la sección 2.1, es posible tener más de una unidad de flujo en una electrofacie, por lo que su lectura no asegura mismas condiciones de capacidad de flujo y/o almacenamiento. Segundo, el algoritmo tiene ciertas limitaciones de dimensionalidad como tamaños y formas uniformes de los *clusters*, por lo que cuando esta condición es violada, el algoritmo puede comportarse de manera intuitiva, arrojando resultandos inconsistentes.

Finalmente, el modelo *Hierarchical Clustering* es usado con los mismos parámetros que el modelo anterior no supervisado. Los resultados se muestran en la tabla 15.

Tabla 15

Reporte de Clasificación Hierarchical Clustering – Agua Grande

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		0.68	0.96	0.79	46
2		0.56	0.10	0.17	49
3		0.32	0.71	0.44	17
	Accuracy			0.54	112
	Micro avg	0.52	0.59	0.47	112
	Weighted avg	0.57	0.54	0.47	112

Nota. Elaboración propia

Como se puede observar de la tabla 15, el modelo presenta los peores valores de exactitud con 54%. Adicionalmente, se comprobó que el modelo falla en el escalamiento de los resultados dado que no cuenta con una función de predicción a partir de una predicción inicial, sino que se adapta siempre a los nuevos datos de entrada, lo cual no genera consistencia en nuevas predicciones a diferencia de los modelos anteriores.

5.2 Caso Sergi-Remanso

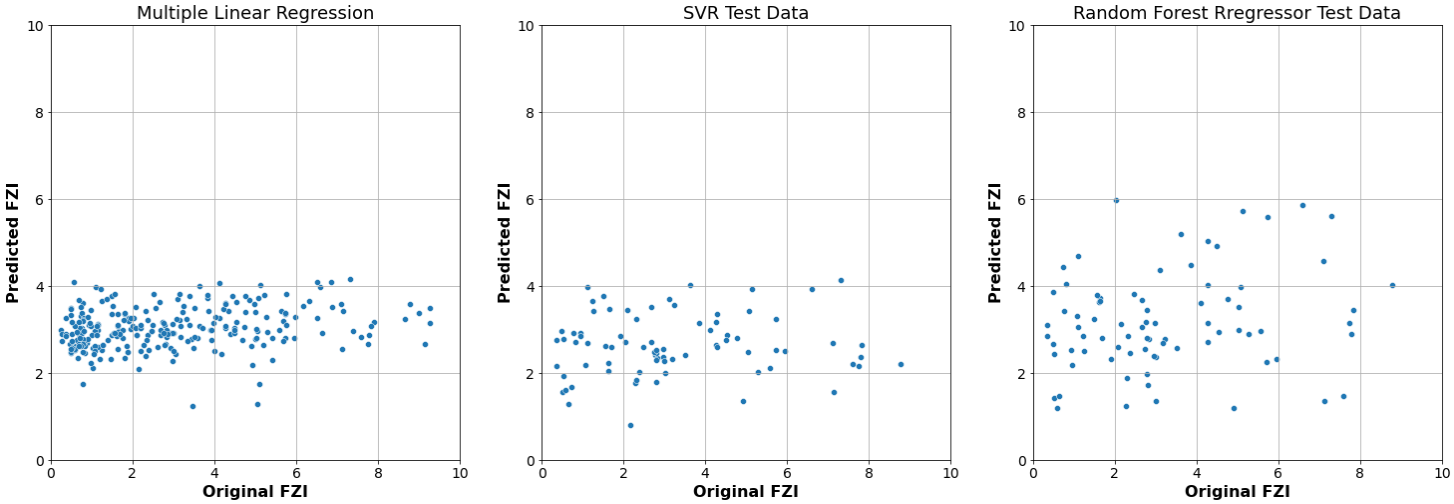
El flujo de trabajo para el caso de estudio Sergi-Remanso es el mismo que para el caso anterior. Se usan los estimadores de regresión y clasificación en el aprendizaje supervisado y los dos algoritmos de *clustering* propuestos para el aprendizaje no supervisado. Tal como se revisó en el análisis exploratorio de datos y de los modelos de porosidad efectiva y permeabilidad al aire determinados, se evaluará la respuesta de los modelos ante la alta heterogeneidad vertical que se encuentra en la formación, con posibles variaciones de tipos de minerales de arcilla producto de la presencia de laminaciones y mineralogía siliciclástica distinta a la observada en el caso Agua Grande-Jandaia producto de la alta respuesta en radioactividad del intervalo mientras los registros densidad y neutrón no manifiestan presencias importantes de lutitas.

Continuando con el flujo de trabajo, se evalúan inicialmente los tres estimadores de regresión: regresión lineal múltiple, *Support Vector Regressor* y *Random Forest Regressor*. A pesar de realizar la búsqueda de los hiperparámetros para los dos últimos estimadores, ambos reportan coeficientes de determinación de 0, al igual que la regresión lineal múltiple, por lo que ninguno de los tres modelos puede ser usado para modelar el FZI en el presente caso de estudio. En la figura 45 se observan los resultados de las tres regresiones.

El primer estimador de clasificación usado, SVM, obtiene hiperparámetros de coeficiente de regularización de 1 y un kernel *rbf*, con una exactitud con validación cruzada del 43%. En la figura 53 se observan las métricas de la evaluación del modelo con el set de prueba. Como se puede observar la tabla 16, el estimador SVM no modela adecuadamente las 4 clases o unidades de flujo identificada, teniendo una exactitud de 43%.

Figura 45

Resultados regresiones caso Sergi



Nota. Elaboración propia.

El siguiente estimador a usar es el *Random Forest Classifier*, se usa el método de estandarización para escalar los datos. La búsqueda de hiperparámetros con el set de entrenamiento y el método de validación cruzada, teniendo en cuenta el peso balanceado de las clases, entrega un número de estimadores óptimo de 5 (árboles de decisión), 3 características máximas a considerar para la división de los nodos de los árboles y 5 muestras mínimas requeridas para estar en un nodo de hoja, con un 43% de exactitud. Al probar el set de evaluación en el modelo, se obtienen los resultados de la tabla 17. Los resultados entregados por el modelo son pobres en comparación con los reportados para el caso de estudio anterior.

Tabla 16

Reporte de Clasificación SVM – Sergi

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		0.38	0.60	0.46	5
2		0.35	0.58	0.44	12
3		0.53	0.38	0.44	21
4		0.50	0.33	0.40	15
	Accuracy			0.43	53
	Micro avg	0.44	0.47	0.44	53
	Weighted avg	0.47	0.43	0.43	53

Nota. Elaboración propia

Tabla 17

Reporte de Clasificación Random Forest Classifier - Sergi

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		0.22	0.33	0.27	6
2		0.14	0.08	0.10	13
3		0.53	0.53	0.53	19
4		0.56	0.67	0.61	15
	Accuracy			0.43	53
	Micro avg	0.36	0.40	0.37	53
	Weighted avg	0.41	0.43	0.41	53

Nota. Elaboración propia

Por último, se usa el estimador *Gradient Boost Classifier,* y el método de estandarización para escalar los datos. Los pesos de las clases presentes son calculados intrínsecamente por el estimador. Se reportan hiperparámetros óptimos de tasa de aprendizaje de 0.1 y 25 estimadores (número de etapas de refuerzo a realizar), con una exactitud de 53% (tabla 18).

Tabla 18

Reporte de Clasificación Gradient Boost Classifier – Sergi

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		0.33	0.17	0.22	6
2		0.56	0.38	0.45	13
3		0.57	0.63	0.60	19
4		0.50	0.67	0.57	15
	Accuracy			0.53	53
	Micro avg	0.49	0.46	0.46	53
	Weighted avg	0.52	0.53	0.51	53

Nota. Elaboración propia

Por los resultados observados en las tablas 16, 17 y 18, se consideran que los modelos de clasificación de aprendizaje supervisado no realizan predicciones consistentes y escalables en comparación con el caso de estudio anterior. Tal como se comentó previamente, la alta heterogeneidad del intervalo analizado puede estar influyendo significativamente en el rendimiento de los modelos. Dado que las unidades de flujo identificadas se obtuvieron de los análisis de núcleos y estos poseen un muestreo mucho más pequeño que las variables de entrada del modelo (registros estandarizados), se concluye que, para obtener unidades de flujo consistentes a partir de técnicas de aprendizaje supervisado en ambientes de alta heterogeneidad vertical, se requieren de variables de entrada con mayor resolución vertical, tales como registros de imágenes de pozo (BHI), por ejemplo. Para el aprendizaje no supervisado, se declaran las variables de entrada como el set de registros estandarizado (GR, CNSS, RHOB) puestos a profundidad con las unidades identificadas tal como para el aprendizaje supervisado. El algoritmo implementado es *K*-

Means¸ sin embargo, dado que de manera previa se conocen los *clusters* por las unidades identificadas, se usan 4 *clusters* en el algoritmo. Los resultados se muestran en la tabla 19.

Tabla 19

Reporte de Clasificación K-Means – Sergi

Class	*Metrics*	*Precision*	*Recall*	*F1-Score*	*Support*
1		0.10	0.40	0.16	25
2		0.37	0.22	0.28	63
3		0.46	0.44	0.45	96
4		0.52	0.20	0.29	80
	Accuracy			0.31	264
	Micro avg	0.36	0.31	0.29	264
	Weighted avg	0.42	0.31	0.33	264

Nota. Elaboración propia

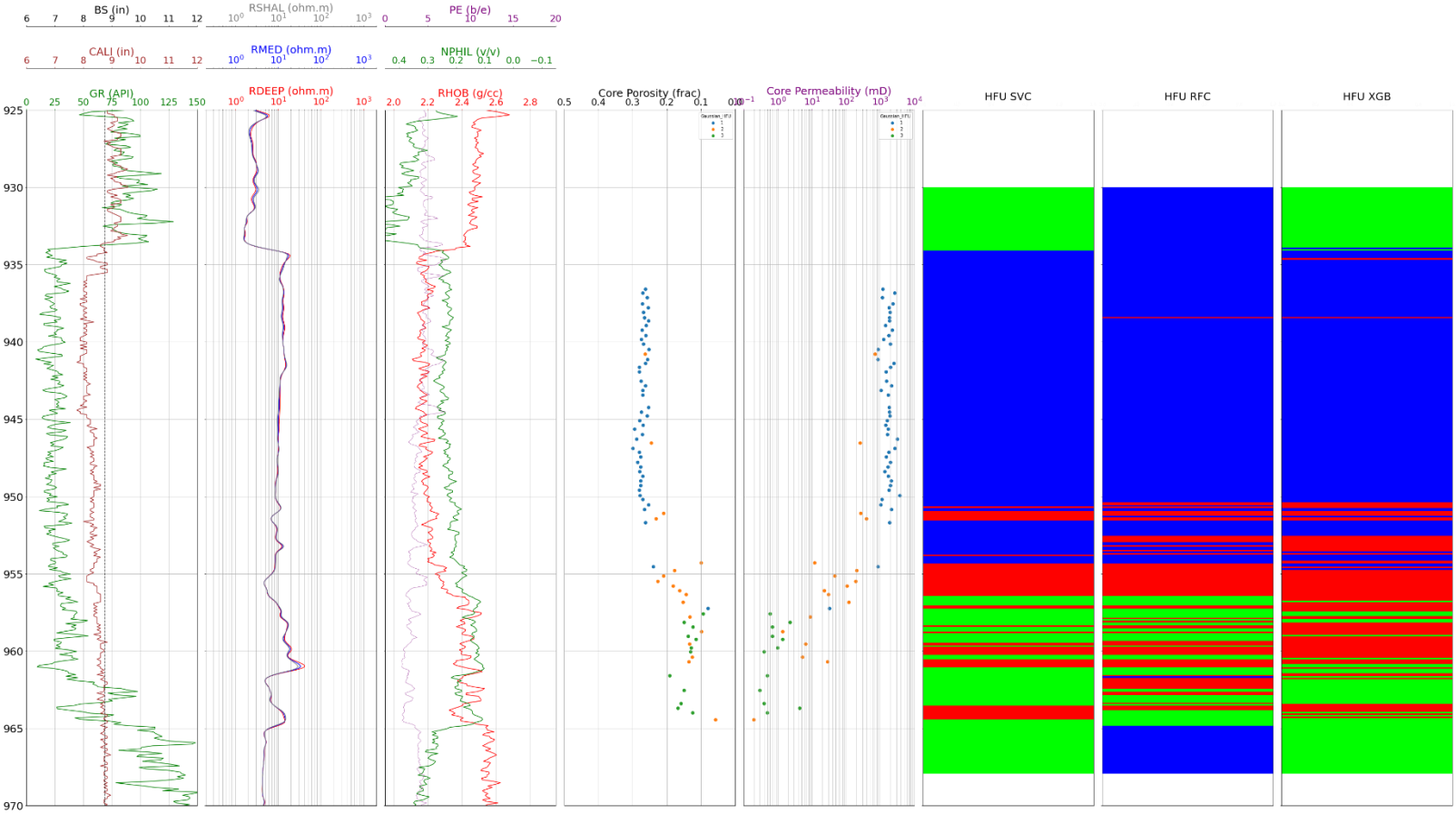
El algoritmo de aprendizaje no supervisado *K-Means* no modela las unidades de flujo identificadas según los resultados de la tabla 19. Sin embargo, en el escalamiento de los resultados en el siguiente capítulo se analizará la posible utilidad de este tipo de algoritmos sobre los supervisados en la identificación de electrofacies en ambientes de alta heterogeneidad vertical.

6. Análisis de Resultados

Primeramente, se escalan los resultados de los mejores modelos de aprendizaje supervisado (*Support Vector Classifier*, *Random Forest Classifier*, *Gradient Boost Classifier*) en el caso de estudio Agua Grande-Jandaia sobre cada uno de los pozos. En la figura 46 se observan para el pozo ***7-JND-3D-BA*** y en la figura 47 para el pozo ***7-JND-13D-BA***.

Figura 46

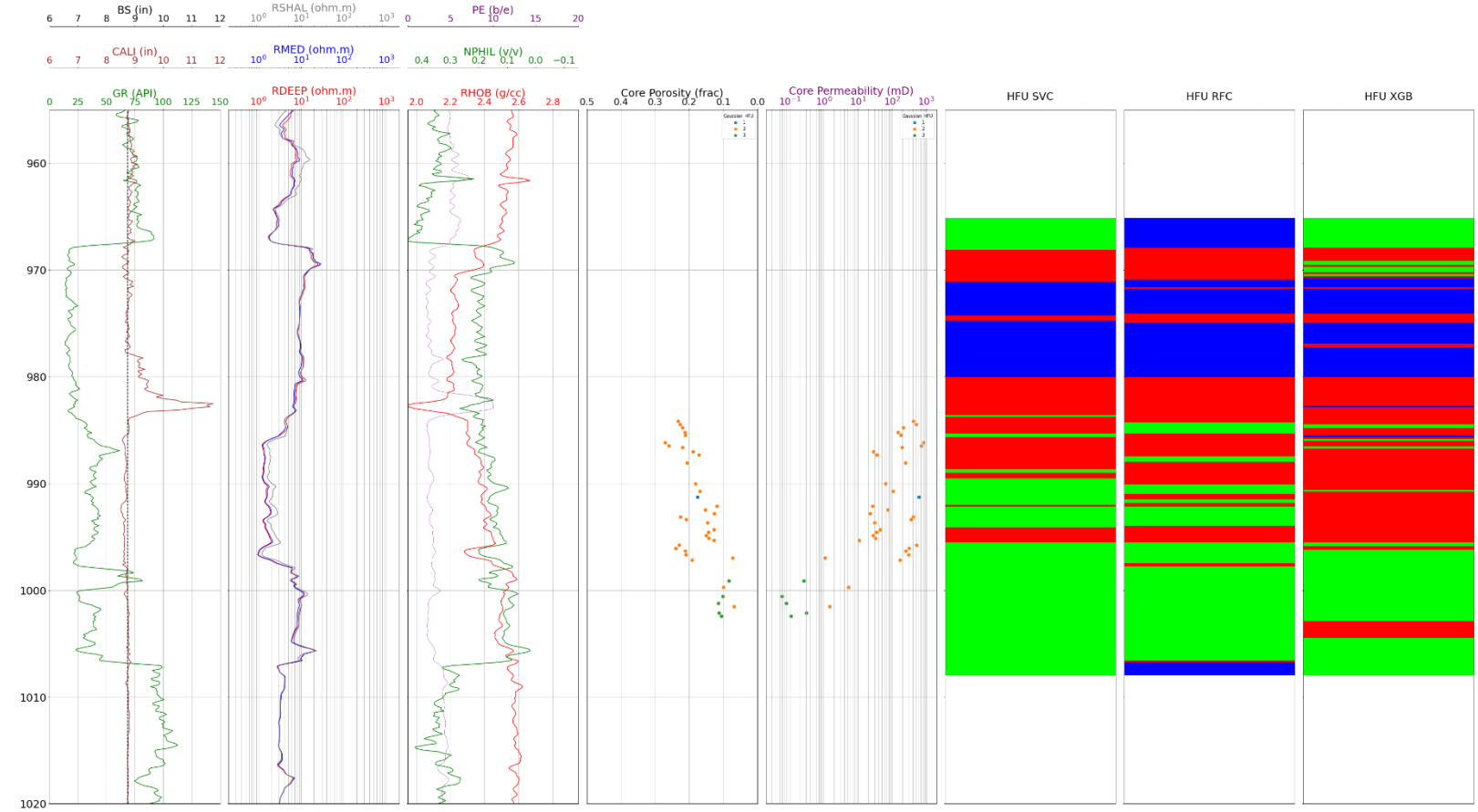
Comparación modelos de Machine Learning supervisados pozo 7-JND-3D-BA



Nota. Elaboración propia en el lenguaje Python. Profundidades en *measured depth* [m].

Figura 47

Comparación modelos de Machine Learning supervisados pozo 7-JND-13D-BA



Nota. Elaboración propia en el lenguaje Python. Profundidades en *measured depth* [m]. Unidad 1 representada por el color azul, unidad 2 por rojo y unidad 3 por verde.

Dado que los modelos fueron entrenados con los registros en profundidad con los núcleos, el escalamiento de cada modelo sobre el intervalo de interés se considera como una predicción sobre datos nuevos nunca vistos en el entrenamiento.

La primera observación importante de las figuras 46 y 47 consiste en la inconsistencia por parte del modelo *Random Forest Classifier* (RFC) en la identificación de la clase o unidad 1. Ya que al extender el intervalo un par de metros arriba y debajo de la formación Agua Grande, identifica intervalos arcillosos como parte de la unidad 1 (color azul), lo cual no es cierto.

Seguidamente, tal como se encontró en el capítulo anterior, el modelo *Support Vector Classifier* (SVC), tiende a clasificar intervalos en unidad 3 (color verde) donde no corresponde. Finalmente, el modelo *Gradiente Boost Classifier* (XGB) representa el mejor ajuste tal como se reportó en los resultados del capítulo anterior. En la figura 46, se observa en la última pista cómo este modelo detalla de forma más precisa la transición entre la unidad 1 y unidad 2 (color naranja), evidenciada en los datos de núcleos. Así mismo, tal como se puede observar de la figura 47, el tope de la formación Agua Grande para el pozo ***7-JND-13D-BA*** no tiene data de núcleo disponible, por lo que la predicción de unidad 1 sobre los registros de este pozo tiene un poco más de incertidumbre, sin embargo, los modelos mantienen la consistencia en la predicción de esta unidad en comparación con la figura 46.

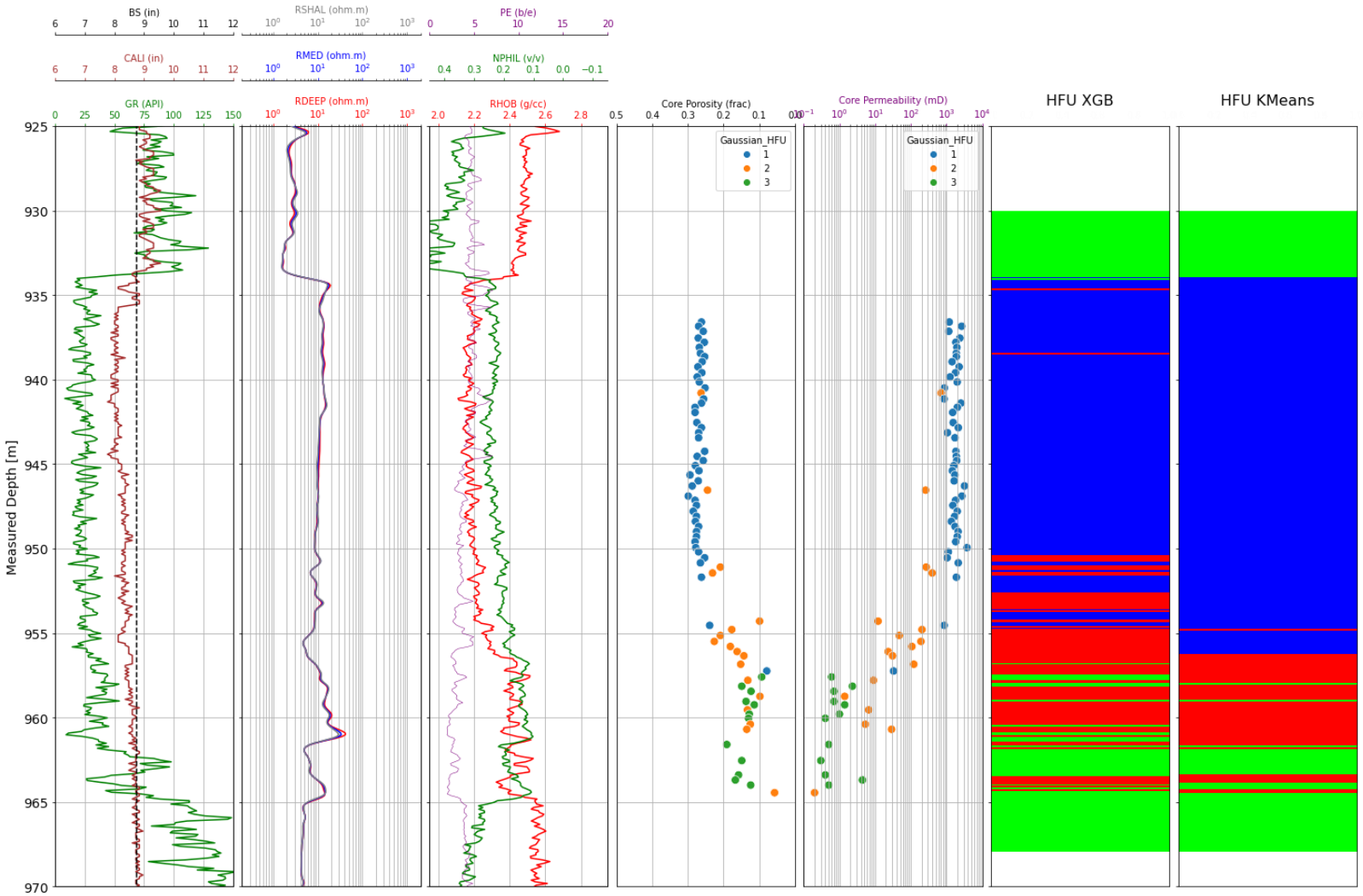
En las figuras 48 y 49 se muestra la comparación entre el mejor modelo de aprendizaje supervisado (*Gradient Boost Classifier*) y no supervisado (*K-Means)* para los dos pozos del caso de estudio Agua Grande-Jandaia.

En el primer pozo (figura 48), se observa una buena consistencia de las unidades de flujo identificadas en los datos de núcleos con las modeladas con los algoritmos presentados. Sin embargo, se puede observar que el algoritmo supervisado *Gradient Boost Classifier* muestra con más detalle la transición entre la unidad 1 y 2 que el algoritmo *K-Means*, el cual responde principalmente a las electrofacies del intervalo bajo el entrenamiento de exactamente los mismos datos que el algoritmo supervisado, teniendo en cuenta que para *K-Means* no se usa la etiqueta de unidades de flujo en el entrenamiento. Así mismo, la alta dispersión obtenida en la identificación de la unidad 3 en la sección 4.1, genera una diferencia menos significativa para el modelo con respecto a la unidad 2, lo que hace que los mayores valores de permeabilidad al aire de la unidad 3 estén asociados por el modelo a la unidad 2, sin embargo, según las propiedades promedio de

cada unidad en la figura 31, la poca calidad de roca contemplada en la unidad 3 no representa motivo para considerar una división con sus valores más cercanos a la unidad 2. Así mismo, la tasa de muestreo entre el núcleo y el registro puede afectar levemente su identificación.

Figura 48

Comparación modelos de Machine Learning pozo 7-JND-3D-BA



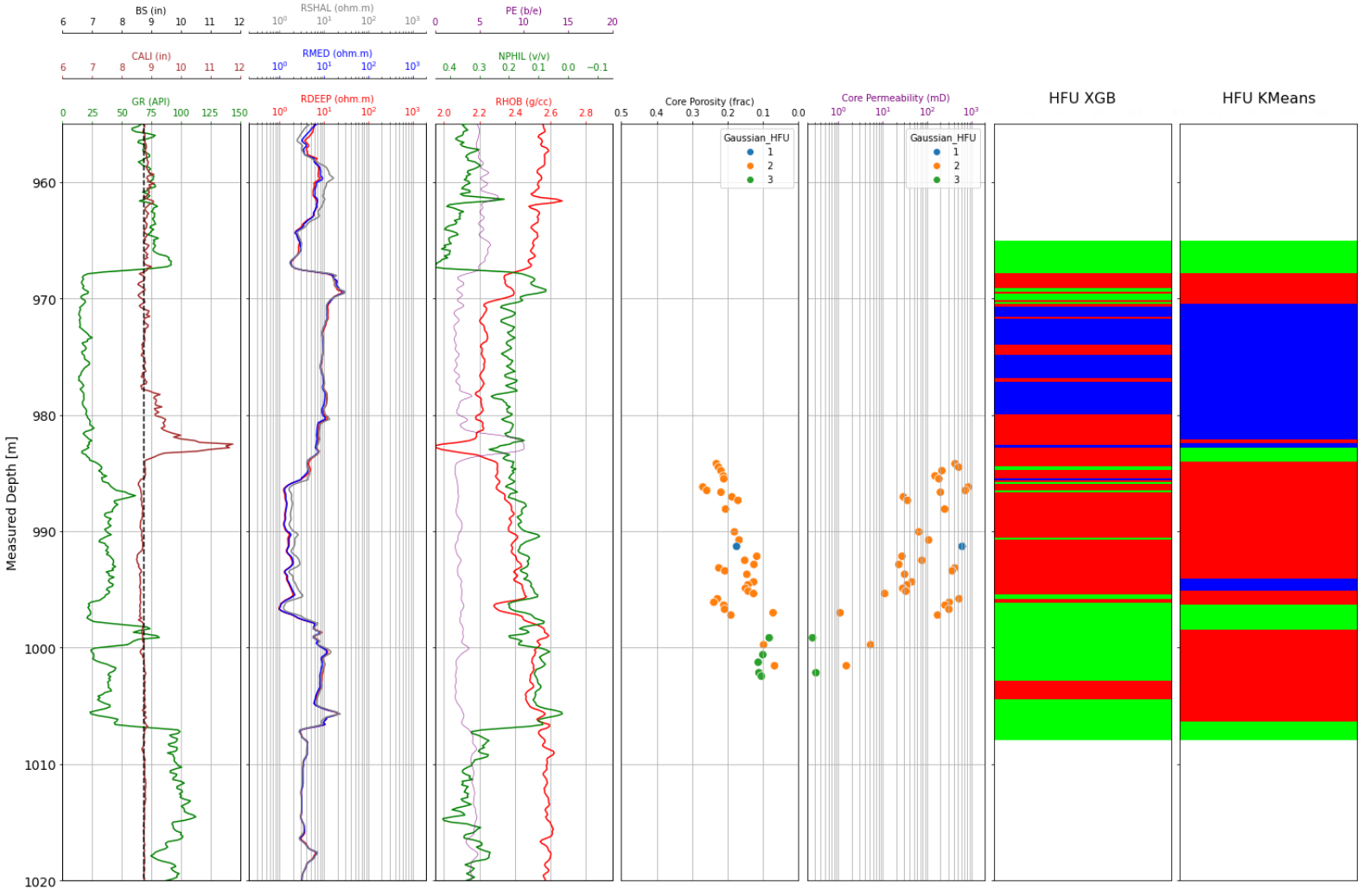
Nota. Elaboración propia en el lenguaje Python. Unidad 1 representada por el color azul, unidad 2 por rojo y unidad 3 por verde.

En el caso del segundo pozo (figura 49), se observa de igual forma consistencia en las unidades identificadas principalmente por el algoritmo supervisado *Gradient Boost Classifier*. El

algoritmo *K-Means* presenta diferencias grandes hacia la base del intervalo, identificando unidad 2 cuando incluso en los datos de núcleo predomina la unidad 3. Así mismo, hacia el tope del intervalo no se aprecia la continuidad de la unidad 1 observada en la figura 48. Para comparar de mejor forma ambos resultados, se realizan *crossplots* de densidad-neutrón con las unidades identificadas por el modelo en la figura 50.

Figura 49

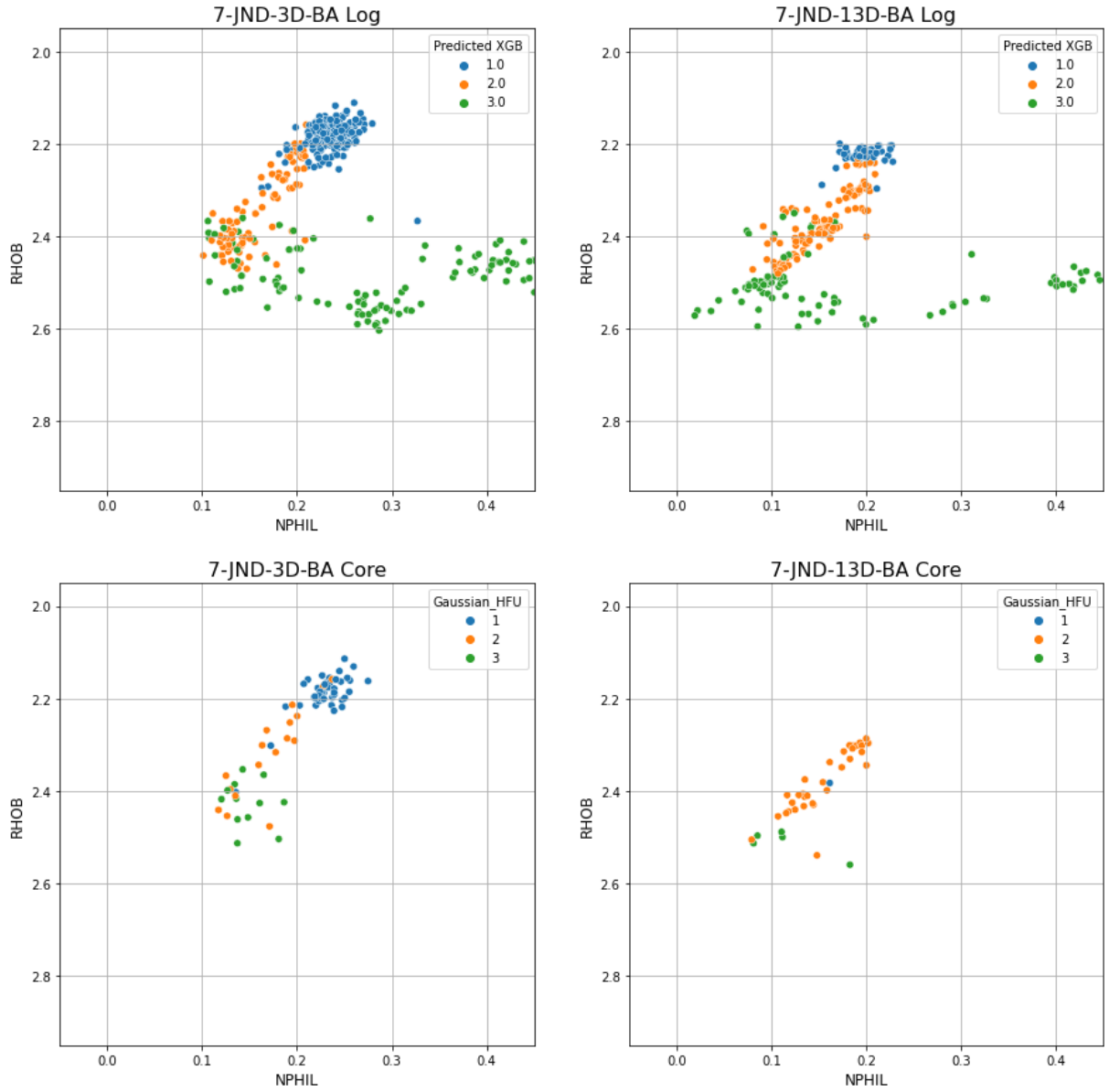
Comparación modelos de Machine Learning pozo 7-JND-13D-BA



Nota. Elaboración propia en el lenguaje Python. Unidad 1 representada por el color azul, unidad 2 por rojo y unidad 3 por verde.

Figura 50

Crossplots densidad-neutrón para Agua Grande según unidad de flujo modelada por XGB



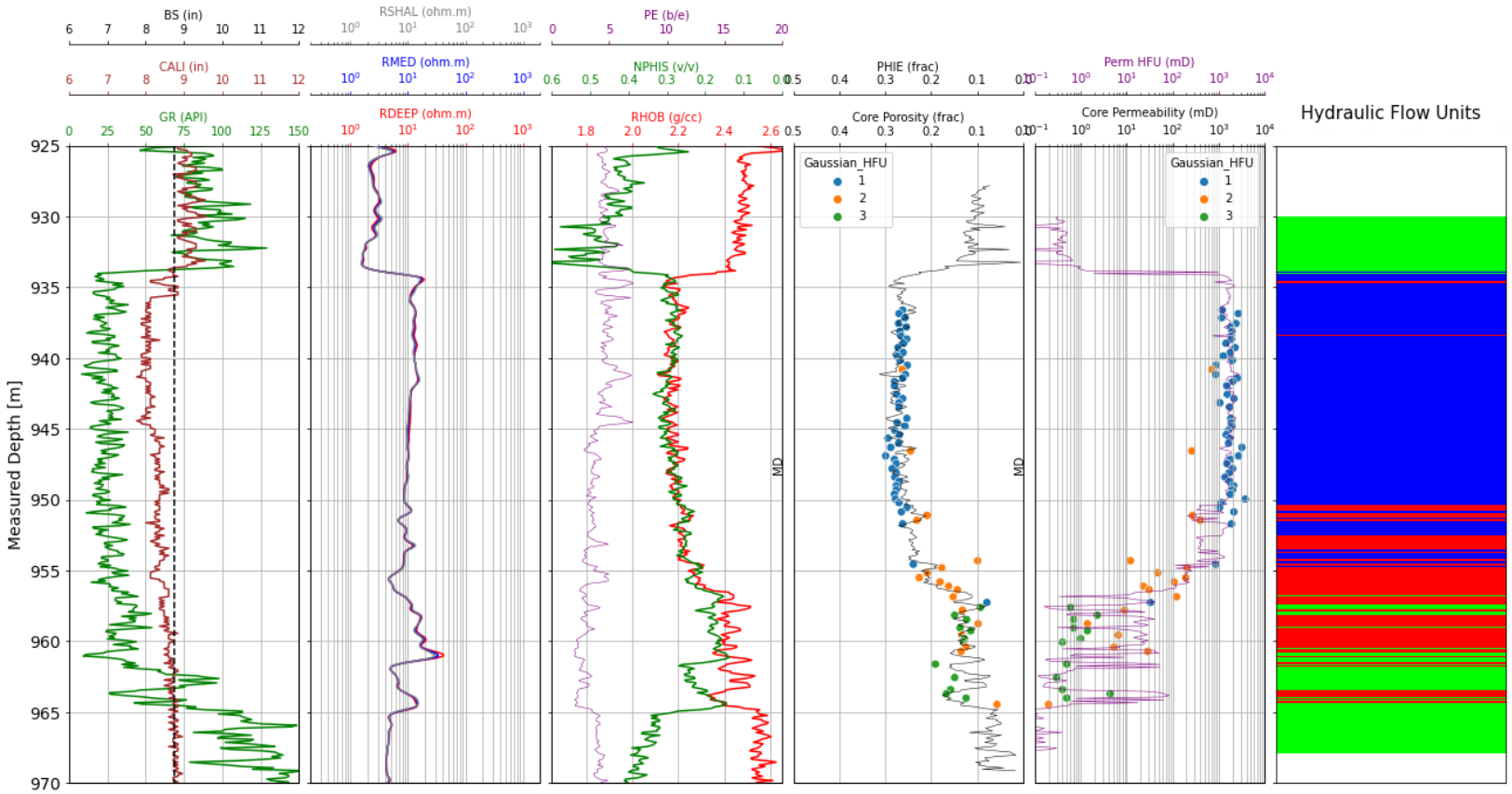
Nota. Elaboración propia.

La figura 50 confirma la consistencia observada en las figuras 48 y 49 de las unidades de flujo modeladas por el algoritmo *Gradient Boost Classifier*, en donde la unidad 1 representa la unidad de mejor calidad de roca y mayores valores de porosidad, seguida de la unidad 2 y por último la unidad 3.

Finalmente, en la figura 51 (neutrón matriz arena) y 52 (neutrón matriz caliza) se presenta el resultado final de las unidades de flujo modeladas con sus respectivos modelos de porosidad efectiva y permeabilidad.

Figura 51

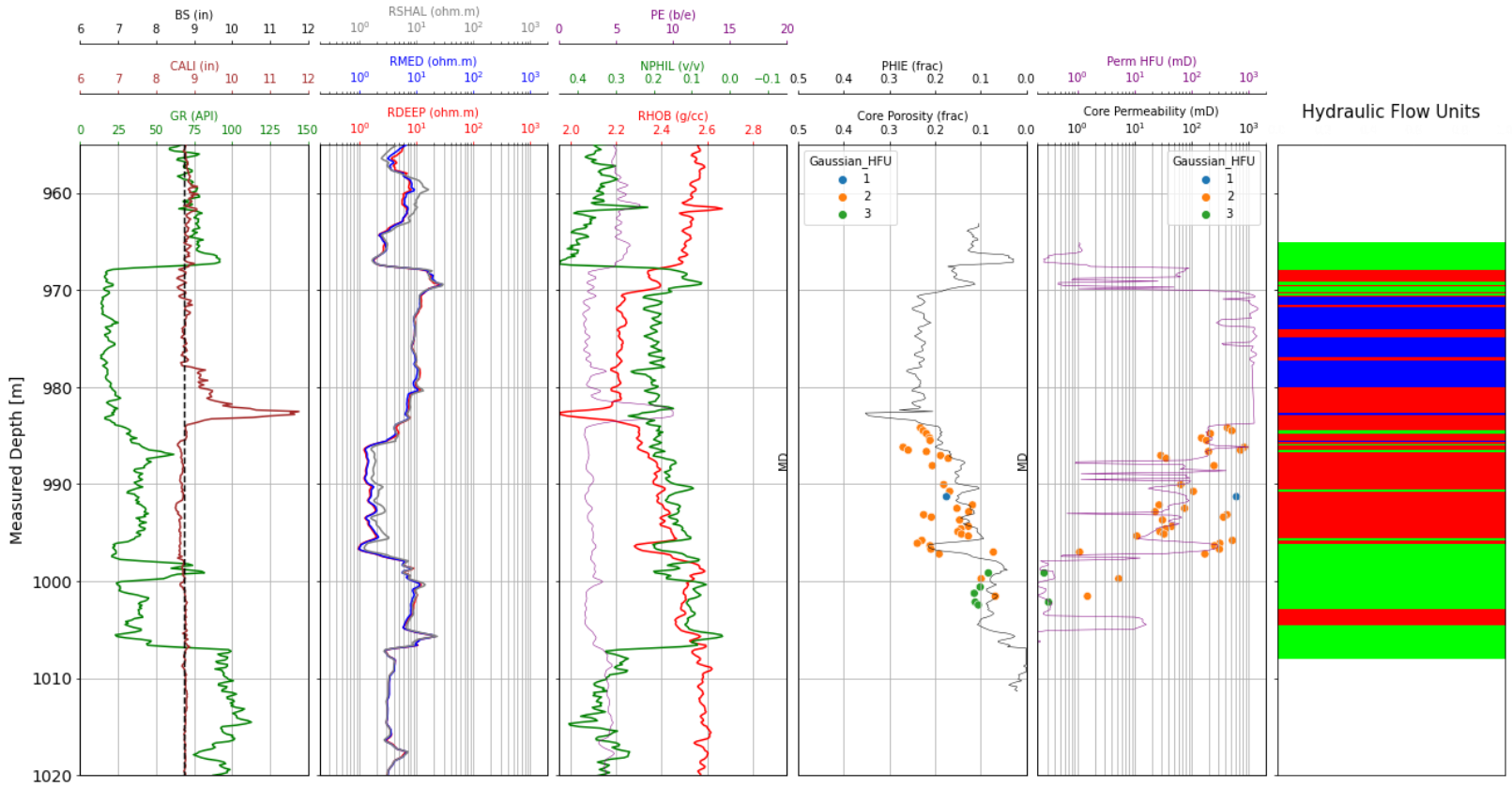
Resultado final pozo 7-JND-3D-BA con unidades de flujo hidráulicas modeladas



Nota. Elaboración propia en el lenguaje Python. Unidad 1 representada por el color azul, unidad 2 por rojo y unidad 3 por verde.

Figura 52

Resultado final pozo 7-JND-13D-BA con unidades de flujo hidráulicas modeladas



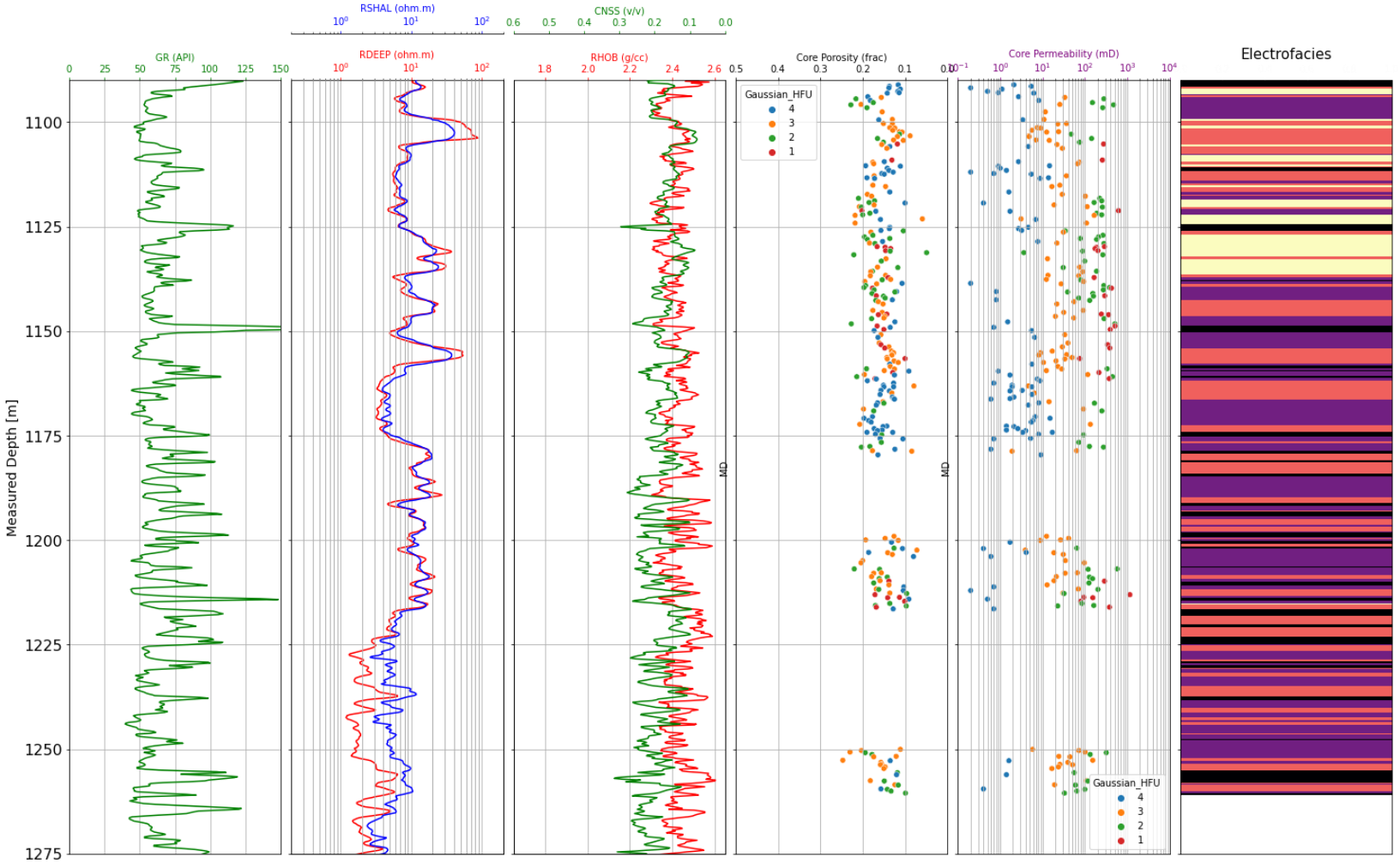
Nota. Elaboración propia en el lenguaje Python. Unidad 1 representada por el color azul, unidad 2 por rojo y unidad 3 por verde. Se observa influencia de zona lavada con baja calidad de hoyo en los resultados del intervalo.

Para el caso de estudio Sergi-Remanso, tal como se concluyó del capítulo anterior, los modelos de aprendizaje supervisado y no supervisado no modelan aceptablemente las unidades de flujo hidráulicas identificadas. Sin embargo, el algoritmo de aprendizaje no supervisado *K-Means* es planteado como una alternativa para la identificación de electrofacies donde el aprendizaje supervisado carece de consistencia. En la figura 53 se observa el resultado final del modelamiento de 4 electrofacies. De forma preliminar, la electrofacie 4 (color amarillo) se observa predominantemente en los *crossovers* entre los registros densidad y neutrón (matriz arena),

indicando intervalos gasíferos posiblemente. Así mismo, la electrofacie 1 (color negro) se observa predominantemente en las zonas de mayor separación en los registros densidad y neutrón y altos valores de GR, indicando zonas de mayor arcillosidad. En la figura 54 se exponen en *crossplots* densidad-neutrón las electrofacies modeladas a partir de los registros y su comparación con las unidades de flujo identificadas de los núcleos.

Figura 53

Resultado final pozo 7-RO-14-BA con electrofacies K-Means

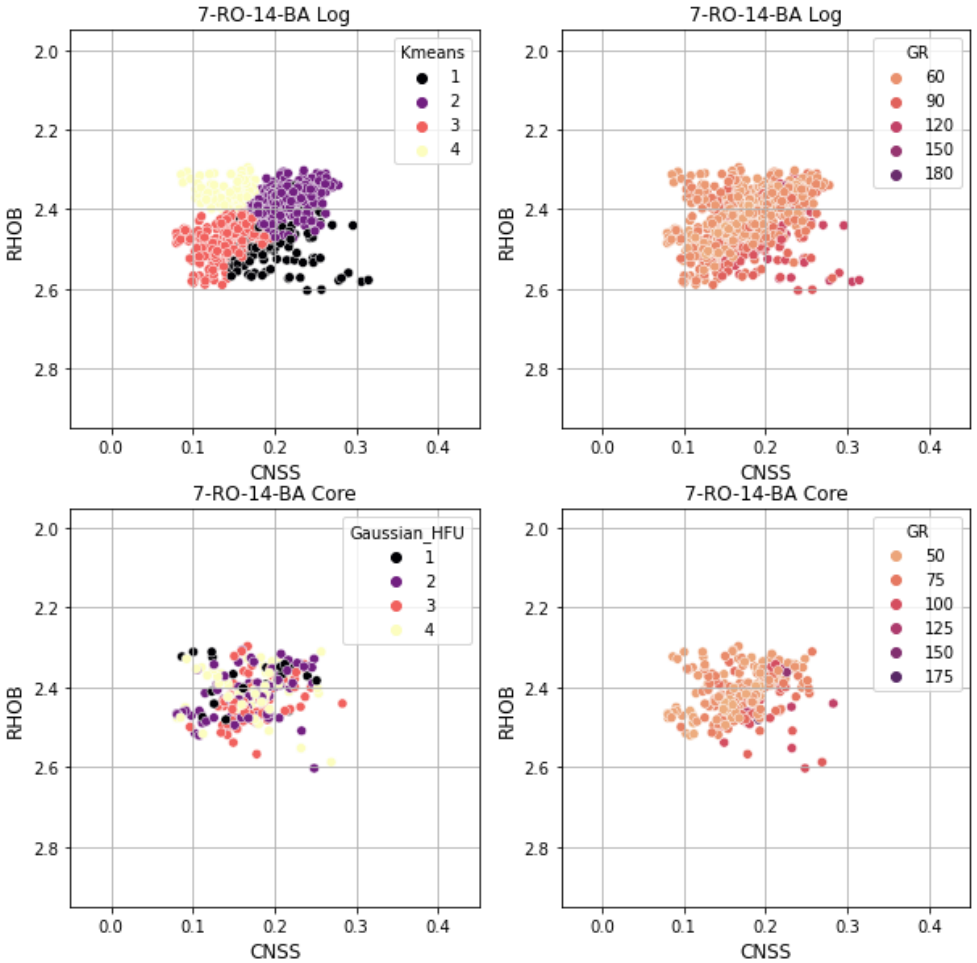


Nota. Elaboración propia en el lenguaje Python. Electrofacie 1 color negro, electrofacie 2 color morado, electrofacie 3 color naranja, electrofacie 4 color amarillo.

Como se puede observar de la figura 54 y en consecuencia con el rendimiento de los estimadores analizados del capítulo anterior, las unidades de flujo (Gaussian_HFU) no se correlacionan con las electrofacies modeladas. Sin embargo, como se puede concluir del *crossplot* y de la representación en la figura 53, los algoritmos de aprendizaje no supervisado tienen la capacidad de modelar electrofacies que faciliten la interpretación de intervalos de alta heterogeneidad vertical donde los registros no alcanzan el muestreo suficiente para ser usados en algoritmos de aprendizaje supervisado con etiquetas de alta resolución como data de núcleos.

Figura 54

Crossplots densidad-neutrón para caso Sergi con Electrofacies K-Means



Nota. Elaboración propia.

Conclusiones

Partiendo de un análisis exploratorio de datos extraídos de la página de archivos técnicos con acceso gratuito de la Agencia Nacional de Petróleo de Brasil, se recopiló la data de análisis rutinario de núcleos en las formaciones Pojuca, Agua Grande y Sergi de la cuenca del Recóncavo, encontrando un comportamiento multimodal en las propiedades petrofísicas medidas en laboratorio, evidenciando la presencia de múltiples condiciones de flujo y almacenamiento. Se plantearon dos casos de estudio según la data disponible y la calidad de la porosidad del núcleo y permeabilidad al aire reportada como un primer caso en Agua Grande y un segundo caso en Sergi. Aplicando el algoritmo de mezclas gaussianas y los métodos del codo y análisis de silueta sobre el conjunto de datos del indicador de zona de flujo propuesto por Amaefule, se identificaron 3 y 4 unidades de flujo en Agua Grande y Sergi, respectivamente. Usando métodos gráficos a partir de los registros GR, Neutrón y Densidad, se establecen *end points* para determinar los modelos de porosidad total y efectiva de cada pozo, aproximando así la porosidad del núcleo medida como la porosidad efectiva. Mediante correlaciones entre la permeabilidad al aire y la porosidad efectiva, se emplean modelos de permeabilidad al aire para cada unidad de flujo identificada.

Al aplicar algoritmos de aprendizaje supervisado y no supervisado para predecir unidades de flujo, tomando como datos de entrada al modelo los registros GR, Neutrón y Densidad, y el indicador de zona de flujo o la unidad de flujo según el modelo, se pudo obtener para un primer caso de estudio con dos pozos de la formación Agua Grande una exactitud de 70% a 80% usando modelos de aprendizaje supervisado, siendo el mejor estimador el *Gradient Boost Classifier*, mientras que los modelos de aprendizaje no supervisado alcanzaron un 60% de exactitud en promedio. Para el segundo caso de estudio con un pozo en la formación Sergi, los algoritmos de

aprendizaje supervisado presentaron menor exactitud que el caso anterior. Esto puede ser ocasionado por la alta heterogeneidad vertical evidenciada en el intervalo, donde los registros de pozo no son capaces de tomar medidas comparables con el muestreo y la resolución de los análisis de núcleos, de los cuales provienen las unidades de flujo verdaderas. Sin embargo, ante estas circunstancias, se desarrolla el uso de algoritmos de aprendizaje no supervisado como el *K-Means* para facilitar la interpretación del intervalo de estudio en la determinación de electrofacies, en donde los algoritmos de aprendizaje supervisado carecen de utilidad y flexibilidad sin la información consistente previa a modelar.

Recomendaciones

Las métricas de los rendimientos en los modelos de *Machine Learning* aplicados en la presente investigación pueden mejorar con la disponibilidad de mayor cantidad de datos, haciendo una base de datos mucho más robusta y permitiendo a los modelos un aprendizaje más consistente. Así mismo, se recomienda la obtención de la permeabilidad Klinkenberg o absoluta para la obtención de los modelos de permeabilidad absoluta en las unidades de flujo identificadas.

Con el fin de correlacionar las unidades de flujo hidráulicas modeladas usando registros eléctricos con las propiedades texturales de la roca, se recomienda la validación con análisis petrográficos, tales como espectroscopías electrónicas de barrido y análisis mineralógicos, así como el uso de datos históricos de productividad para los intervalos en estudio.

Por último, debe tenerse en cuenta que la comparación de múltiples de registros de pozo requiere de un procesamiento adecuado dado que las herramientas realizan corridas con diferentes configuraciones de diseño y calibración. Por ende, la normalización de los registros por intervalo de interés es altamente recomendada previamente a la aplicación de cualquier técnica interpretativa multipozo.

Anexos

Los archivos de *Jupyter Notebook* usados en la presente investigación bajo el lenguaje Python se pueden encontrar en el siguiente link bajo la licencia allí descrita:

https://github.com/jeasierraan13/ML-HFU.git

Referencias Bibliográficas

Abbas, Mohammed A, and Erfan M Al Lawe. "Clustering Analysis and Flow Zone Indicator for Electrofacies Characterization in the Upper Shale Member in Luhais Oil Field, Southern Iraq." Paper presented at the Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, UAE, November 2019. doi: https://doi-org.bibliotecavirtual.uis.edu.co/10.2118/197906-MS

Amaefule, Jude O., Altunbay, Mehmet, Tiab, Djebbar, Kersey, David G., and Dare K. Keelan. "Enhanced Reservoir Description: Using Core and Log Data to Identify Hydraulic (Flow) Units and Predict Permeability in Uncored Intervals/Wells." Paper presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, October 1993. doi: https://doi.org/10.2118/26436-MS

Baldwin, J. L., Bateman, R. M., & Wheatley, C. L. (1990). Application of a neural network to the problem of mineral identification from well logs. The Log Analyst, 31(05), 279–293.

Busch, J., Fortney, W., & Berry, L. J. S. (1987). Determination of lithology from well logs by statistical analysis. SPE Formation Evaluation, 2(04), 412–418

Centro de Investigaciones en Geofísica y Geología (CPGG-UFBA). (2008). Estudo dos sistemas petrolíferos das bacias do recôncavo, tucano e jatobá. Universidad Federal de Bahía

Daya Sagar, Cheng, Q., & Agterberg, F. (2018). *Handbook of Mathematical Geosciences*. Springer International Publishing AG.

Delfiner, P., Peyret, O., & Serra, O. J. S. (1987). Automatic determination of lithology from well logs. SPE Formation Evaluation, 2(03), 303–310.

Fadokun, Daniel Oluwadara, Oshilike, Ishioma Bridget, and Mike Obi Onyekonwu. (2020) Supervised and Unsupervised Machine Learning Approach in Facies Prediction. Paper presented at the SPE Nigeria Annual International Conference and Exhibition, Virtual. doi: https://doi.org/10.2118/203726-MS

Fazel Alavi, M. "Determination of Reservoir Permeability Based on Irreducible Water Saturation and Porosity from Log Data and FZI (Flow Zone Indicator) from Core Data." Paper presented at the International Petroleum Technology Conference, Doha, Qatar, January 2014. doi: https://doi.org/10.2523/IPTC-17429-MS

Gill, D., Shomrony, A., & Fligelman, H. J. A. B. (1993). Numerical zonation of log suites and logfacies recognition by multivariate clustering. AAPG Bulletin, 77(10), 1781–1791.

Glover, P. W. J. (2010). Petrophysics. Department of Geology and Petroleum Geology. University of Aberdeen. UK.

Hong, Youngjun, Wang, Shinjo, Bae, Jeehoon, Yoo, Jaeyoon, and Sungroh Yoon. "Automated Facies Identification Using Unsupervised Clustering." Paper presented at the Offshore Technology Conference, Houston, Texas, USA, May 2020. doi: https://doi-org.bibliotecavirtual.uis.edu.co/10.4043/30773-MS

Kadkhodaie-Ilkhchi R, Rezaee R, Moussavi-Harami R, Kadkhodaie -Ilkhchi A. Analysis of the reservoir electrofacies in the framework of hydraulic flow units in the Whicher Range Field, Perth Basin, Western Australia. Journal of petroleum science & engineering. 2013;111:106–20

Kapur, L., Lake, L. W., Sepehrnoori, K., Herrick, D. C., & Kalkomey, C. T. (1998) Facies prediction from core and log data using artificial neural network technology. In SPWLA 39th annual logging symposium, 1998. Society of Petrophysicists and Well-Log Analysts.

Khalid, M., Desouky, SD., Rashed, M. et al. (2020). Application of hydraulic flow units' approach for improving reservoir characterization and predicting permeability. J Petrol Explor Prod Technol 10, 467–479. https://doi.org/10.1007/s13202-019-00758-7

Magnavita, L. Szatmari, P. Cupertino, J. Destro, N. Roberts, D. (2012). 15 – The Reconcavo Basin. Regional Geology and Tectonics: Phanerozoic Rift Systems and Sedimentary Basins. Pages 382-419. Elsevier. https://doi.org/10.1016/B978-0-444-56356-9.00014-6.

Mello, M. R., Koutsoukos, E. A. M., Mohriak, W. U., Bacoccoli, G. (1994). Selected Petroleum Systems in Brazil. AAPG.

Narayan, I. Rastogi, A. Kainkaryam, S. Bhattacharya, S. Saputelli, L. (2020). Machine Learning in the Oil and Gas Industry. Springer Science. https://doi.org/10.1007/978-1-4842-6094-4

Prates, I. (2017). Cuenca del Recóncavo: Resumen Geológico y sectores en oferta. Superintendencia de Definición de Bloques, Brasil.

Robail, Frederic, Sanyal, Satyashis, B M Noor Azudin, Ahmad Nazmi, Koh, Kwi Yen, Bt Hairon Nizar, Farahani, and Ummi Farah Mohamad Rosli. (2023). Machine Learning for Facies Distribution of Large Carbonate Reservoir Models- A Case Study. Paper presented at the International Petroleum Technology Conference, Bangkok, Thailand, March 2023. doi: https://doi.org/10.2523/IPTC-22876-MS

Rogers, S. J., Fang, J., Karr, C., & Stanley, D. J. A. (1992). Determination of lithology from well logs using a neural network. AAPG Bulletin, 76(5), 731–739

Shi, Xinlei, Chen, Hongbing, Li, Ruijuan, Yang, Xiaoyan, Liu, Huan, and Ting Li. (2019) Improving Permeability and Productivity Estimation with Electrofacies Classification and Core Data Collected in Multiple Oilfields. Paper presented at the Offshore Technology Conference, Houston, Texas. doi: https://doi.org/10.4043/29214-MS