Minería de datos para el análisis de la continuidad de las propiedades del reservorio en las escamas Veloz 1, 2 y 3 del yacimiento Yumurí-Seboruco, Cuba

Ariel García Martínez[1], Odalys Reyes-Paredes[2] y Milton García Borroto[3]

[1] *Ingeniero geofísico. Centro de Investigación del Petróleo, DIGICUPET. Calle 23 No. 105 e/ O y P, Plaza. C.P. 10400. La Habana Cuba. Correo electrónico: agarciam@digi.cupet.cu.*

[2] *Ingeniera geofísica. Máster en Geofísica Aplicada, Mención Petróleo y Gas. Centro de Investigación del Petróleo, DIGICUPET. Calle 23 No. 105 entre O y P, Plaza. C.P. 10400. La Habana Cuba. Correo electrónico oreyes@digi.cupet.cu.*

[3] *Licenciado en Ciencias de la Computación. Doctor en Ciencias de la Computación. Universidad Tecnológica de la Habana «José Antonio Echeverría», CUJAE, ave 114 No. 11901 entre Ciclovía y Rotonda, Marianao, CP: 19390, La Habana, Cuba. Correo electrónico: mgarciab@ceis.cujae.edu.cu*

RESUMEN

En el yacimiento Yumurí-Seboruco perteneciente a la Franja Petrolífera Norte Cubana, no se logra identificar zonas perspectivas en las escamas Veloz 1, 2 y 3, por lo que el objetivo de la presente la investigación es evaluar la continuidad de las propiedades del reservorio para lograr la identificación, mediante técnicas de minería de datos. Para ello se realizó la descripción de la imagen sísmica en las zonas conocidas como reservorios enfocada a la identificación de cambios en el comportamiento del campo ondulatorio que justificasen la aplicación de las técnicas de minería de datos, a través del uso de atributos sísmicos. Los datos aportados por los volúmenes de atributos sísmicos fueron muestreados y transformados en puntos para facilitar su procesamiento en la confección de bases de datos de entrenamiento y predicción. Con las bases de datos confeccionadas se pasó a entrenar las técnicas de minería de datos seleccionadas (redes neuronales, árbol de decisión J48 y *RandomForest*) y a la construcción de modelos para cada una de las características del reservorio estudiadas (saturación de agua, fracturación y calidad del reservorio). Se compararon los resultados alcanzados por cada algoritmo en el entrenamiento y se relacionaron con la información disponible en los pozos y escoger el algoritmo que predice los modelos que más se ajustan a la geología. Con los modelos seleccionados, se pasó a caracterizar la continuidad en el comportamiento de cada una de las características en la zona correspondiente al reservorio, se identificaron rasgos tectónicos que delimitan el comportamiento de dichas características y se identificaron zonas perspectivas con la presencia de hidrocarburos en el sector.

Palabras clave: árboles de decisión, atributos sísmicos, minería de datos, redes neuronales, sísmica.

ABSTRACT

In the Yumurí-Seboruco deposit belonging to the Northern Cuban Oil Belt, it is not possible to identify prospective zones in the Veloz 1, 2 and 3 scales, so the objective of this research is to evaluate the continuity of the reservoir properties to achieve identification, through data mining techniques. For this, the descrip-

García-Martínez, A., Reyes-Paredes, O. y García-Borroto, M. 2021, Minería de datos para el análisis de la continuidad de las propiedades del reservorio en las escamas Veloz 1, 2 y 3 del yacimiento Yumurí-Seboruco, Cuba: Geociencias UO. v. 4, núm. 1, enero-junio 2021. pp. 52-64

tion of the seismic image was made in the areas known as reservoirs focused on the identification of changes in the wave field behavior that justify the application of data mining techniques, through the use of seismic attributes. The data provided by the volumes of seismic attributes were sampled and transformed into points to facilitate their processing for the preparation of training and prediction databases. With the ready-made databases, the selected data mining techniques (neural networks, decision tree J48 and RandomForest) were trained, and the construction of models for each of the reservoir characteristics studied (water saturation, fracturing and reservoir quality). The results achieved by each algorithm in the training were compared and related to the information available in the wells to choose the algorithm that predicts the models that best fit the geology. With the selected models, we continued to characterize the continuity in the behavior of each of the characteristics in the area corresponding to the reservoir, tectonic features that were delimiting the behavior of these characteristics were identified and prospective areas for the presence of hydrocarbons in the sector.

Keywords: decision trees, seismic attributes, data mining, neural networks, seismic.

RESUMO

No campo Yumurí-Seboruco pertencente ao Cinturão Cubano do Norte, não é possível identificar zonas prospectivas nas escalas Veloz 1, 2 e 3, portanto o objetivo desta pesquisa é avaliar a continuidade das propriedades do reservatório para obter a identificação , por meio de técnicas de mineração de dados. Foi feita a descrição da imagem sísmica nas áreas conhecidas como reservatórios, focadas na identificação de mudanças no comportamento do campo de ondas que justificam a aplicação de técnicas de mineração de dados, através do uso de atributos sísmicos. Os dados fornecidos pelos volumes de atributos sísmicos foram amostrados e transformados em pontos para facilitar seu processamento para a preparação de bancos de dados de treinamento e previsão. Com os bancos de dados prontos, foram treinadas as técnicas de mineração de dados selecionadas (redes neurais, árvore de decisão J48 e RandomForest) e a construção de modelos para cada uma das características do reservatório estudadas (saturação da água, faturamento e qualidade do reservatório). Os resultados alcançados por cada algoritmo no treinamento foram comparados e relacionados às informações disponíveis nos poços para escolher o algoritmo que prediz os modelos que melhor se ajustam à geología. Com os modelos selecionados, continuamos a caracterizar a continuidade no comportamento de cada uma das características na área correspondente ao reservatório, foram identificadas características tectônicas que delimitavam o comportamento dessas características e áreas prospectivas para a presença de hidrocarbonetos na região ou setor.

Palavras-chave: árvores de decisão, atributos sísmicos, mineração de dados, redes neurais, sísmicas.

INTRODUCCIÓN

Las modernas técnicas de minería de datos permiten a los especialistas extraer información de grandes volúmenes de datos y determinar características que serían imposibles detectar en estudios convencionales. El uso de este tipo de técnicas dentro de los estudios geofísicos ha cobrado relevancia a partir de todos los avances que se han dado dentro de la informática. En el caso de los estudios sísmicos petroleros son muy utilizadas para apoyar la interpretación en zonas de geología compleja, mediante la búsqueda de relaciones entre distintas variables y la localización de áreas con características similares. Este es el caso del yacimiento Yumurí-Seboruco perteneciente a la Franja Petrolífera Norte Cubana (FPNC), donde existe un ambiente geológico compresivo de cinturones plegados y cabalgados que traen como consecuencia un alto grado de inclinación de las fronteras, al limitar el papel de la sísmica en la caracterización de las estructuras existentes y la localización de nuevos prospectos. Los reservorios más importantes del país, localizados en la FPNC, donde se origina la mayor parte de la producción de crudo en Cuba se encuentran en rocas carbonatadas. El yaci-

miento Yumurí–Seboruco, se ubica al norte de la provincia de Matanzas, a una distancia aproximada de 75 km de la ciudad de La Habana, al comprender un área total de desarrollo de unos 33 km^2. (**Figura 1**).

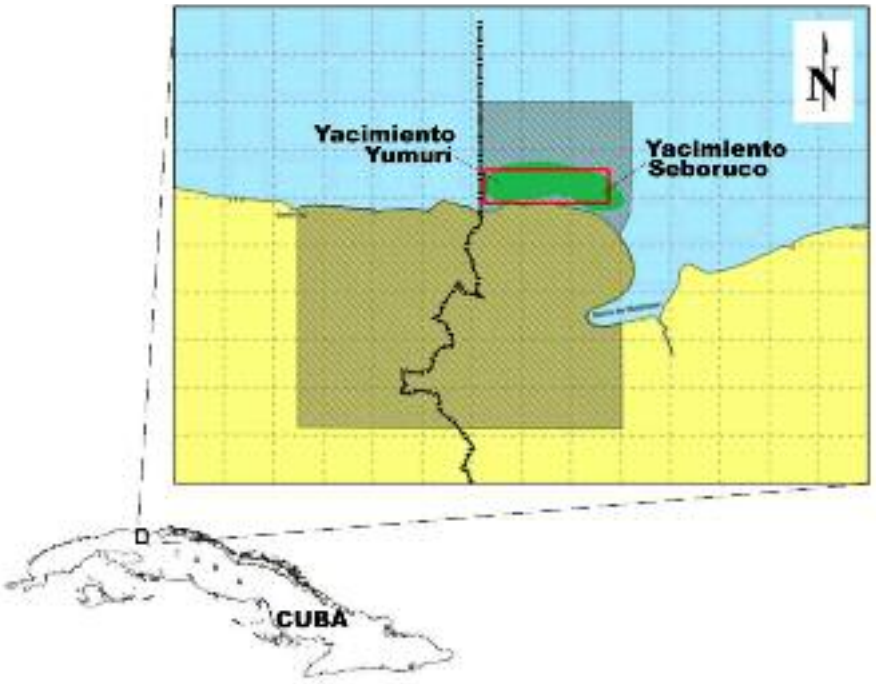

Figura 1. Ubicación del yacimiento Yumurí–Seboruco. Norte de la provincia de Matanzas.

Muchos autores han realizado investigaciones relacionadas con el uso de métodos de minería de datos para la detección de patrones sísmicos en combinación con los atributos sísmicos y la información que aportan los registros geofísicos de pozos, apoyándose en técnicas que van desde las Redes Neuronales Artificiales (RNA) y los mapas auto-organizados, hasta las máquinas de soporte vectorial y los árboles de decisión, al resaltar en todos los casos, que ninguna técnica de clasificación es superior a otra, sus capacidades para obtener mejores resultados dependen de la naturaleza de los datos y la característica que se desea clasificar (Zhao *et al*., 2015).

En el caso específico de Cuba, los trabajos realizados se han enfocado en lo fundamental a la detección de fluidos para la búsqueda directa de hidrocarburos, la estimación de registros sintéticos de pozos y análisis de similitud para pequeñas zonas dentro del volumen sísmico, destacándose el proyecto para la formación como ingeniero de Fernández Mederos (2016) y la tesis en opción al título de Master en Ciencias Geofísicas de Villavicencio García (2005), en la que se estima un registro sónico sintético a partir de los registros geofísicos disponibles en un pozo. La investigación de Gómez Herrera (2019), realiza la caracterización de los intervalos potencialmente productores de gas en la FPNC, mediante la aplicación de un análisis exploratorio de datos y métodos de minería de datos, al lograr identificar los intervalos litofaciales patrones con sus firmas sísmicas.

A pesar de su complejidad, el área de estudio del sector Yumurí–Seboruco presenta un elevado interés para el desarrollo energético del país, debido a sus potencialidades para la presencia de hidrocarburos. En este contexto, en la zona se han realizado estudios integrales a partir de datos geológicos, geofísicos, geomorfológicos y geoquímicos en la parte marina del sector Yumurí–Seboruco. Es objetivo de este trabajo, mejorar el modelo geológico del yacimiento para el análisis de la continuidad de las pro-

piedades del reservorio, la detección de nuevas áreas perspectivas, la perforación de nuevos pozos y la caracterización de cada una de las escamas que componen el reservorio carbonatado fracturado.

MATERIALES Y MÉTODOS.

Para el desarrollo de este trabajo, se utilizó un volumen sísmico 3D migrado en profundidad, a partir del cual se construyeron los volúmenes de atributos sísmicos. Se utilizaron cinco pozos para la construcción de la base de datos de entrenamiento y 18 pozos para la validación de los modelos predichos, además se trabajó con los registros geofísicos de pozos disponibles.

Los métodos empleados para lograr los resultados de esta investigación fueron geofísicos (métodos sísmicos y de pozo) y estadísticos matemáticos, valiéndose de técnicas como: atributos sísmicos y la minería de datos.

Selección de atributos

Para realizar el entrenamiento se tomaron los pozos Seboruco (Seb) A, Seb-B, Seb-C, Seb-D y Seb-D2, debido a que fueron interpretados como parte de los estudios realizados en el área pertenecientes al Proyecto 9030 del Ceinpet «Evaluación integral de prospectos para la exploración petrolera en el sector Yumurí-Seboruco» (**Tabla 1**). En este proyecto se determina para cada uno de los 5 pozos el comportamiento que posee la fracturación, la saturación de agua (sw) y la calidad del reservorio, dentro de las escamas V1, V2 y V3.

Debido a la facilidad que implica tener ese grupo de pozos ya interpretados, se tomaron esas tres características como las que debían ser tenidas en cuenta por los modelos para realizar su entrenamiento.

El siuiente paso en el análisis de la información existente, fue el muestreo de los atributos para

Pozo	Manto	Φe [%]	Sa [%]	Ksist [mD]	λ [adim]	r_{35} [µm]	k/Φ [mD]	Calidad	Observaciones
Seb-A	V1	14	22	2734	4	0.34	20169	B	Fracturado
Seb-L	V. Gray	16	35	4075	4	0.33	26734	B	Muy Fracturado
	V. Seb.	18	32	4568	4	0.60	27594	B	Muy Fracturado
	V. Blue	12	43	2747	4	0.19	25645	I	Fracturado, Sw media
Seb-N	V. Blue	13	79	1092	5	0.09	7460	P	Poco fracturado, Sw alta
	V1	14	31	2036	3	0.44	13835	B	Fracturado
Seb-B	V. Blue	11	50	4164	4	0.15	35881	P	Muy fracturado, Sw alta
	V1	12	24	4110	3	0.48	29632	B	Muy Fracturado
	V2	15	16	4999	2	1.03	38352	B	Muy Fracturado y matriz conductiva
Seb-C	V1	14	45	853	3	0.49	6156	I	Poco fracturado, Sw media
	V2	13	24	1025	3	0.28	7680	I	Poco fracturado
	V3	17	16	1626	4	0.45	11622	B	Fracturado
Seb-D	V. Blue	13	48	1842	3	0.46	31801	I	Fracturado, Sw media
	V1	14	33	2389	3	0.39	27201	B	Fracturado
Seb-D2	V. Blue	12	56	6419	4	0.24	59037	P	Muy fracturado, Sw alta
	V1	12	52	5117	3	0.26	37901	P	Muy fracturado, Sw alta
	V2	13	43	7262	3	0.38	61051	B	Muy fracturado

B (Buena), I (Intermedia), P (Pobre)

Tabla 1. Caracterización de los pozos seleccionados para el entrenamiento, tomada de (Delgado Ramos *et al*., 2019).

asociarlos a cada uno de los pozos de entrenamiento. Este muestreo se realizó en el programa de interpretación Petrel al transformar toda la información de los atributos correspondiente a cada una de las escamas, en puntos con coordenadas X, Y y Z, con un paso de muestreo de 10 m. De todos los puntos resultantes solo se tuvieron en cuenta para el análisis los que se encontraban en un radio de 50 m alrededor de cada uno de los 5 pozos seleccionados.

Para realizar la selección de los atributos, el entrenamiento y la construcción de los modelos se empleó el programa informático WEKA (*Waikato Enviroment for Knowledge Analysis*- Ambiente para el Análisis del Conocimiento de la Universidad de Waikato). WEKA es un programa potente de minería de datos, compuesto por una serie de herramientas graficas de visualización y diferentes algoritmos para el análisis de datos y modelos predictivos (Calleja Gómez, 2010).

Los datos son cargados en WEKA, se selecciona la pestaña *attribute selection* (selección de atributos), al permitir explorar que subconjunto de atributos son los que mejor pueden clasificar la característica estudiada. Como método de búsqueda se usó el Ranker, el cual se encarga de ordenar los atributos según alguna medida de correlación, esta medida se fija en el método de evaluación (Corso, 2010), que para este caso fue seleccionado el InfoGainAttributeEval. **La Tabla 2** muestra el resultado de la selección de atributos para la saturación de agua y la calidad del reservorio en la escama 3, debido a que para estas dos características la selección de atributos es la misma. Se tomó la escama Veloz 3 como ejemplo porque es el resultado de la combinación de las bases de datos de las escamas Veloz 1 y Veloz 2, se destaca como le otorga el mayor peso a la información aportada por las velocidades y la componente de bajas frecuencias, lo que concuerda con la interpretación hecha de los atributos sísmicos. El dato sísmico y la amplitud original aparecen con el mismo peso, lo que indica que están muy correlacionados, al final de la tabla aparecen los atributos amplitud RMS y suavizado estructural.

En el caso de la fracturación (**Tabla 3**) la información aportada por la velocidad se mantiene como la de mayor importancia, en segundo y tercer lugar aparecen la frecuencia dominante y la atenuación, lo que concuerda con la interpretación de los atributos sísmicos en la que se identificó que la frecuencia dominante estaba al responder a la fracturación y no a la presencia de fluidos. El dato sísmico y la amplitud original mantienen el mismo peso en la clasificación, al indicar una correlación fuerte y se mantienen en último lugar la amplitud RMS y el suavizado estructural. Para complementar la información aportada por la selección de atributos se construyó una matriz de correlación con la información de todos los datos disponibles hasta el momento (**Tabla 4**). La matriz confirma que el dato sísmico (variable 9) presenta una correlación de 1 con la amplitud original (variable 10), la am-

Media	Atributo
0.870	**Velocidad**
0.296	Componente de bajas frecuencias
0.062	Calidad Instantánea
0.058	Intensidad de reflexión
0.047	Atenuación
0.040	Componente de medias frecuencias
0.038	Dato sísmico
0.038	Amplitud original
0.029	Frecuencia dominante
0.020	Curvatura
0.017	Impedancia acústica relativa
0.012	Componentes de altas frecuencias
0.008	**Amplitud RMS**
0.001	**Suavizado estructural**

Tabla 2. Selección de atributos para la calidad del reservorio y la saturación de agua en la escama Veloz 3.

Media	Atributo
0.805	**Velocidad**
0.215	Frecuencia dominante
0.209	Atenuación
0.200	Dato sísmico
0.200	Amplitud original
0.189	Curvatura
0.155	Componentes de bajas frecuencias
0.135	Calidad instantánea
0.081	Intensidad de la reflexión
0.079	Componentes de medias frecuencias
0.029	Impedancia acústica relativa
0.028	Componentes de altas frecuencias
0.0015	**Amplitud RMS**
0.0012	**Suavizado estructural**

Tabla 3. Selección de atributos para la fracturación en la escama Veloz 3.

Variable	Media	Desviación Estándar
1	-0.006	2.2558
2	0.004	0.0319
3	1.274	0.5843
4	0.011	0.0057
5	0.000	1.0398
6	5.003	2.6732
7	0.150	0.3445
8	0.046	0.0225
9	0.002	1.4057
10	0.002	1.4057
N1	3870.1	172.5
N2	0.949	0.5819
N3	0.755	0.6238
N4	0.544	0.4552

(a)

V*	1	2	3	4	5	6	7	8	9	10	N1	N2	N3	N4
1	1													
2	-0.027	1												
3	0.007	-0.098	1											
4	0.002	0.015	-0.287	1										
5	0.123	-0.039	0.062	-0.002	1									
6	0.015	-0.890	**0.945**	-0.291	0.061	1								
7	-0.005	-0.019	-0.176	0.068	0.006	-0.155	1							
8	0.016	-0.535	-0.244	0.661	-0.001	-0.213	0.078	1						
9	0.146	-0.036	0.051	-0.001	**0.934**	-0.052	0.001	0.004	1					
10	0.146	-0.036	0.051	-0.001	**0.934**	0.052	0.001	0.004	**1.000**	1				
N1	0.029	0.185	0.045	-0.070	-0.032	0.037	-0.201	-0.114	-0.028	-0.03	1			
N2	0.244	0.010	0.548	-0.240	0.001	0.515	-0.154	-0.340	-0.004	-0.004	0.144	1		
N3	0.556	-0.219	0.430	-0.102	0.009	0.455	-0.094	-0.030	0.008	0.008	0.009	0.365	1	
N4	0.689	-0.170	0.310	0.023	0.015	0.305	-0.068	0.132	0.018	0.018	-0.015	0.113	0.716	1

V* (Variables) (b)

Tabla 4. Valores de los dos primeros momentos de las variables (a) y la matriz de correlación de los datos seleccionados para el entrenamiento.

plitud RMS (variable 5) presenta una correlación alta con las variables 9 y 10, el suavizado estructural (variable 3) presenta una correlación alta con la calidad instantánea (variable 6). Para corregir esta situación se determinó eliminar estas tres variables (amplitud original, amplitud RMS y suavizado estructural).

Con la eliminación de las variables que menos información aportaban, se confeccionó una base de datos para cada escama, en el caso específico de la escama Veloz 3 debido a la falta de información se decidió que su base de datos fuera una combinación de los datos de las escamas Veloz 2 y Veloz 1, al quedar conformadas de esta forma las bases de datos de entrenamiento por un total de 11 atributos (frecuencia dominante, atenuación, curvatura, calidad instantánea, intensidad de la reflexión, impedancia acústica relativa, el dato sísmico migrado en profundidad, las velocidades y las componentes de altas, bajas y medias frecuencias, la base de datos de V1 posee un total de 43 544 puntos por atributos, la base de datos de V2 12 325 puntos y la base de datos de V3 por 55 869 puntos.

Entrenamiento de los datos y construcción de los modelos de predicción.

La selección de estos algoritmos se fundamenta en la información aportada por la búsqueda bibliográfica en la que se recomienda el uso de técnicas de clasificación supervisada.

Algoritmo J48 y Redes Neuronales.

El algoritmo J48 forma parte del conjunto de algoritmos basados en los árboles de decisión. La peculiari-

dad de este algoritmo se fundamenta a que incorpora una poda del árbol de clasificación una vez que este ha sido inducido, el eliminar de esta forma las ramas del árbol con menor capacidad predictiva (García Jiménez *et al*., 2007).

Para realizar el entrenamiento de este modelo, se cargó la base de datos con extensión csv en el programa WEKA y con la herramienta (*classify*) se selecciona el clasificador J48. Como se planteó, el parámetro más importante a configurar encargado de controlar la poda del árbol, denominado factor de confianza se mantuvo en 0.25 (predeterminado). El entrenamiento para cada uno de los parámetros con una validación cruzada de 10 grupos aportó los siguientes resultados:

La **Tabla 5** muestra los resultados del entrenamiento del algoritmo J48 para las escamas V1, V2 y V3. Se destaca en este caso, que las instancias clasificadas de forma correcta superaron el 90 % de efectividad en todos los casos y el error medio cuadrático se mantuvo bajo en todos los ejemplos. La propiedad mejor clasificada que atiende a estos aspectos es la saturación de agua (SW) y la escama Veloz 2 la que mejor calidad presentó. La calidad del reservorio se representa por las siglas (CR) y la fracturación (FR).

Las redes neuronales constituyen una de las formas de emular el funcionamiento y las características del cerebro humano para asociar hechos y memorizarlos (Catalina, 1994), de modo específico en esta investigación se usó una red de tipo perceptrón multicapa, que aprovecha la naturaleza paralela de las redes neuronales para reducir el tiempo requerido por un procesador secuencial para determinar la salida adecuada a partir de una entrada. Como función de activación se usó la tangente hiperbólica (predeterminada por el programa WEKA) y para realizar el entrenamiento se empleó el algoritmo de retropropagación, que se encarga de propagar el error hacia atrás durante el entrenamiento, es decir, de la capa de salida hacia la capa de entrada, al pasar por las capas ocultas (Castro, 2006).

Escama	Atributo	Clasificaciones		Error Medio Cuadrático
		Correctas [%]	Incorrectas [%]	
Veloz 1	**SW**	**93.28**	**6.72**	**0.052**
	FR	93.15	6.85	0.053
	CR	93.27	6.73	0.052
Veloz 2	SW	98.36	1.64	0.101
	FR	96.31	3.69	0.151
	CR	**96.36**	**3.64**	**0.150**
Veloz 3	**SW**	**93.91**	**6.09**	**0.193**
	FR	92.53	7.47	0.213
	CR	93.35	6.65	0.202

Tabla 5. Calidad de la clasificación del árbol de decisión J48.

El perceptrón multicapa que propone el programa WEKA trae ya implementadas una serie de capas ocultas (a, m, n, i) y se activan de modo individual o se combinan, en dependencia de la eficiencia que se desee lograr. Para este caso se logró el mejor resultado al utilizar la capa i y crear de esta forma una red neuronal formada por 11 neuronas en la capa de entrada que se corresponden con cada uno de los atributos de la base de datos de entrenamiento, 11 neuronas en la capa oculta y tres neuronas en la capa de salida que se corresponden con las clases de la característica que se está clasifica (**Figura 2**).

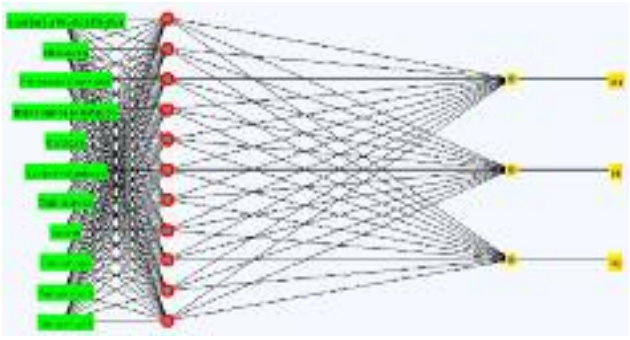

Figura 2. Estructura de la red neuronal usada.

Los resultados alcanzados en el entrenamiento de la red neuronal en cada una de las escamas están resumidos en la **Tabla 6**. El porciento de instancias clasificadas de forma correcta para la red neuronal fue el peor de los tres clasificadores usados, aunque se tiene en cuenta la complejidad geológica de la zona donde se está trabajando, este resultado se mantiene considerado como aceptable.

El entrenamiento para cada uno de los parámetros se realizó con una validación cruzada de 10 grupos, aportó los siguientes resultados:

Escama	Atributo	Clasificaciones Correctas [%]	Clasificaciones Incorrectas [%]	Error Medio Cuadrático
Veloz 1	SW	**88.31**	**11.69**	**0.245**
	FR	87.28	12.72	0.255
	CR	88.30	11.70	0.248
Veloz 2	SW	91.39	8.61	0.206
	FR	**93.78**	**6.22**	**0.178**
	CR	87.89	12.11	0.245
Veloz 3	SW	**84.15**	**15.85**	**0.286**
	FR	78.09	21.91	0.334
	CR	84.10	15.90	0.282

Tabla 6. Calidad de la clasificación para el algoritmo *RandomForest*.

Escama	Atributo	Clasificaciones Correctas [%]	Clasificaciones Incorrectas [%]	Error Medio Cuadrático
Veloz 1	SW	**96.47**	**3.53**	**0.144**
	FR	96.28	3.72	0.146
	CR	96.53	3.47	0.144
Veloz 2	SW	97.46	2.54	0.060
	FR	98.23	1.77	0.111
	CR	**98.36**	**1.64**	**0.111**
Veloz 3	SW	**96.54**	**3.46**	**0.141**
	FR	96.40	3.60	0.151
	CR	96.50	3.50	0.147

Tabla 7. Control de la calidad en la clasificación del algoritmo *RandomForest* para la escama Veloz 1.

Algoritmo *RandomForest.*

RandomForest, también conocido como bosques aleatorios, está formado por combinaciones de árboles predictores de forma tal que cada árbol depende de valores de un vector aleatorio probado de forma independiente y con la misma distribución para cada uno de estos (García Jiménez *et al*., 2007). El entrenamiento para cada uno de los parámetros con una validación cruzada de 10 grupos aportó los siguientes resultados, mostrados en la **Tabla 6**:

El porcentaje de clasificaciones correctas hechas por este algoritmo es superior al de los dos anteriores, debido a que alcanza un 96.97 % como promedio para las tres escamas. La escama que mejor fue clasificada por el modelo fue Veloz 2, al lograr un 98.01 % de clasificaciones correctas como promedio. Las clasificaciones incorrectas no pasan del 5 % en ninguno de los casos y el error medio cuadrático se mantiene con valores más pequeños que con los algoritmos anteriores.

Como este algoritmo mostró resultados superiores a los anteriores fue analizado con más detalle que los anteriores. La **Tabla 7** muestra para la escama Veloz 1 como el TP Rate (*True Positive Rate*) toma valores muy altos para casi todas las clases, al destacar la saturación de agua, en donde la clase saturación alta se logra identificar el 100 % de los elementos que pertenecen de modo real a esa clase. Los falsos positivos (FP) no sobrepasan el 3 % en ninguno de los casos, al demostrar la superioridad del *RandomForest* respecto a los dos algoritmos usados anteriormente. La precisión se mantiene muy alta para todas las clases, al obtener un 98.34 % como promedio. Las tablas correspondientes a las dos escamas restantes se encuentran en los anexos 11 y 12.

Mediante la interpretación de las matrices de confusión representadas en la **Figura 3**, se observa como el modelo comete pocos errores en el proceso de diferenciación de las clases correspondientes a cada una de las categorías. Para la calidad del reservorio las principales equivocaciones son entre las clases buena y pobre, pero si se tiene en cuenta la cantidad que, si son bien clasificadas, estas equivocaciones no afectan la calidad del modelo. Para la fracturación y la saturación de agua el modelo posee un comportamiento similar a la calidad del reservorio en cuanto a que no comete muchas equivocaciones.

Debido a que los resultados alcanzados por el algoritmo *RandomForest* fueron superiores a los dos métodos anteriores, se determinó que sería utilizado para realizar la predicción de las características estudiadas.

RESULTADOS Y DISCUSIÓN

Predicción y validación de los modelos.

Para realizar la predicción se remuestrearon los datos con un paso de 5 metros, al garantizar de esta forma que la base de datos tuviese un tamaño de paso menor al de la usada para el entrenamiento y se cambió la extensión de la base de datos, de extensión csv a arf, debido a que la segunda es la extensión original del programa WEKA.

Para realizar la validación geológica de los datos se contaba con los pozos ya interpretados, de los

Calidad del reservorio

	BN	IN	PB	
Veloz 1	23214	23	159	BN
	85	12909	0	IN
	1140	98	5783	PB
Veloz 2	6821	167	5	BN
	139	5129	0	IN
	3	5	7013	PB
Veloz 3	30040	175	174	BN
	276	17986	0	IN
	1220	105	5696	PB

Fracturación

	FR	MF	PF	
Veloz 1	20765	194	4	FR
	1212	8113	129	MF
	30	44	12920	PF
Veloz 2	118	25	4	FR
	1	13691	175	MF
	1	136	5131	PF
Veloz 3	20855	242	13	FR
	1226	14780	294	MF
	42	188	18032	PF

Saturación de agua

	BA	MD	AL	
Veloz 1	23200	26	170	BA
	88	12905	1	MD
	1145	100	5776	AL
Veloz 2	9566	29	0	BA
	30	2634	2	MD
	0	1	7020	AL
Veloz 3	32710	78	203	BA
	228	15427	5	MD
	1289	122	5610	AL

Figura 3. Matrices de confusión para el control de la clasificación del algoritmo. *RandomForest*.

cuales se obtuvo la saturación de agua y el volumen de arcilla. En el caso de la fracturación fueron utilizados los análisis de densidad de fracturas hechos a partir de los registros de imágenes de microresistividad de formación (FMI).

La **Figura 4** muestra la validación geológica para el modelo predicho por el algoritmo *RandomForest*. Este clasificador solo cometió un error en la estimación de la saturación de agua en el pozo (SEB-F) para las escamas Veloz 2 y Veloz 3, al obtener un 94.2 % de efectividad. Este resultado mejora el obtenido por los clasificadores anteriores y unido al alcanzado por este clasificador en el entrenamiento y la validación matemática, hace que el modelo propuesto por el algoritmo *RandomForest* sea el idóneo para realizar la identificación de nuevas áreas perspectivas para la presencia de hidrocarburos en las escamas de la zona en estudio.

Análisis de los modelos.

Desde el punto de vista de la fracturación el modelo predicho se caracteriza por un dominio de los puntos pertenecientes a la categoría de muy fracturado. Los

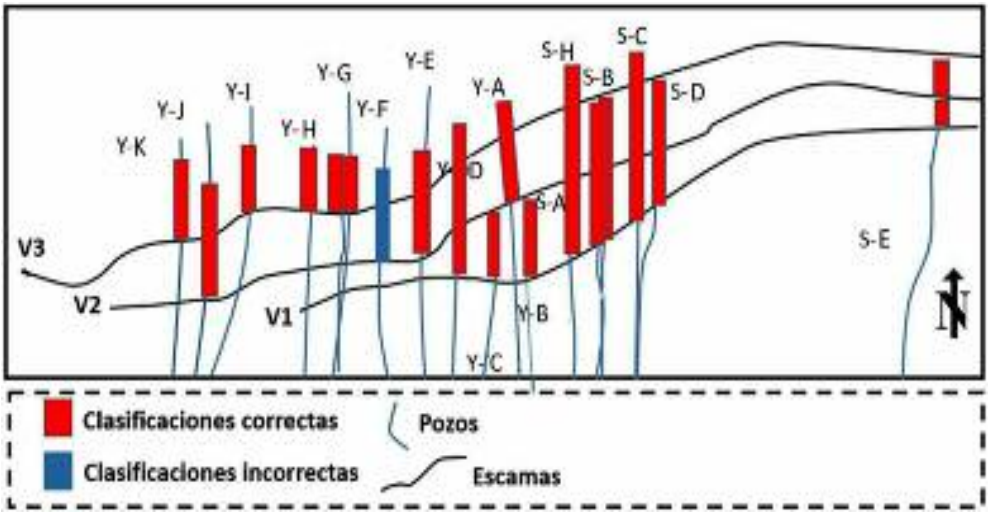

Figura 4. Validación geológica para el modelo predicho por el algoritmo *RandomForest*.

valores pertenecientes a la clase Fracturado se encuentran distribuidos en el área correspondiente al tope de las estructuras, en la parte central del modelo estos valores van desde el tope hasta loa parte inferior del modelo. Los puntos asociados a la categoría Poco Fracturado son los que están representados en menor cantidad (esto era lo esperado debido a que el trabajo se realiza en el área correspondiente a un reservorio fracturado), estos valores aparecen concentrados en la parte central del modelo. Como se observa en la **Figura 5** el modelo muestra un comportamiento en el Este, que se ve interrumpido por una franja de valores muy fracturados a partir de la cual comienza un comportamiento distinto que concentra estos valores en el tope de la estructura. Esta franja de valores fracturados parece indicar la presencia de un elemento tectónico (posible falla) que divide al modelo en dos sectores. Esta diferenciación de sectores es menos evidente para Veloz 3.

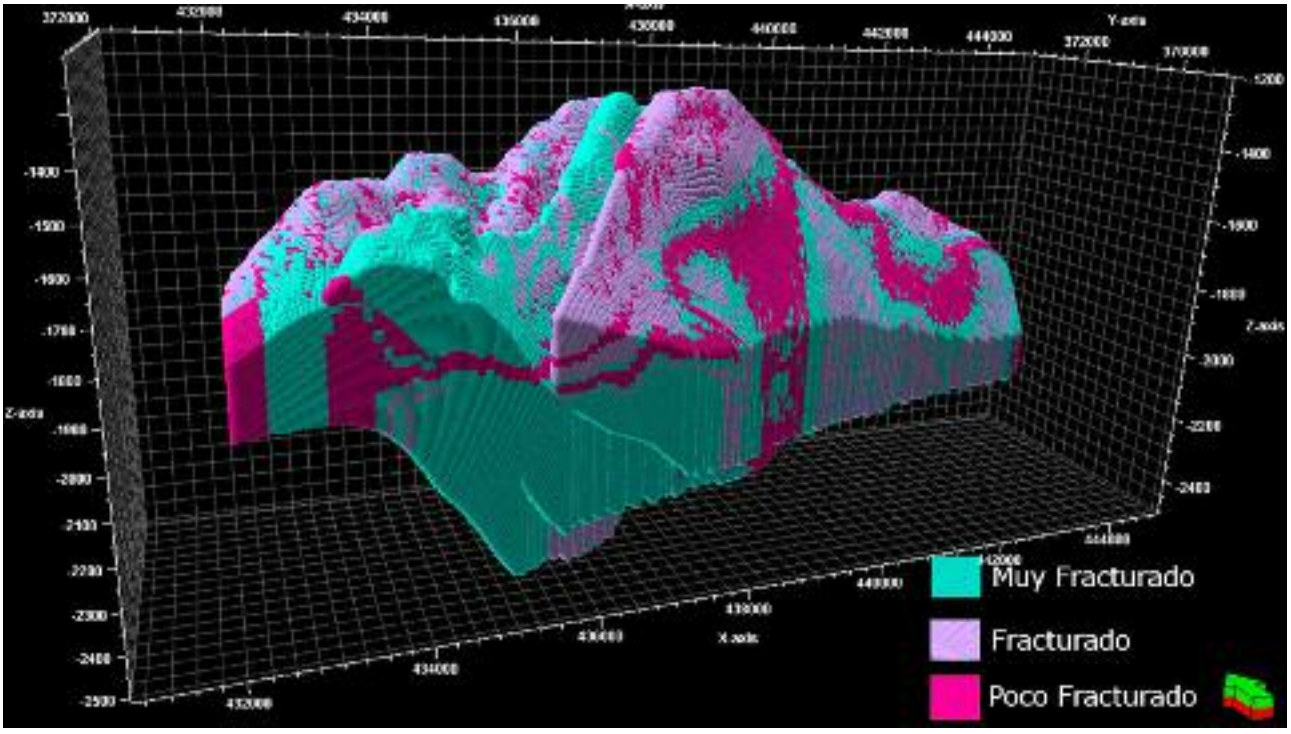

Figura 5. Modelo de fracturación estimado por el algoritmo *RandomForest.*

En el caso de la saturación de agua en la **Figura 6** se observa cómo la categoría predominante está asociada a la clase Baja. Para este modelo aparece de forma continuada una franja en la parte central (en este caso saturada de agua), que divide a la escama en dos zonas con un comportamiento distinto. En la parte Este se aprecia como la clase correspondiente a la saturación de agua baja se distribuye desde el tope de la estructura hasta la parte inferior del modelo, los valores de saturación media son mínimos y la saturación alta se ubica en un conjunto de pequeñas zonas en la parte inferior. Al oeste del contacto el modelo presenta un comportamiento distinto, debido a que las zonas con saturación baja parecen limitarse al tope de la estructura, aumentan los valores de saturación media y se concentran en la parte central, cerca de la posible falla. Los valores de saturación de agua alta ocupan el menor número de puntos y aparecen en la parte inferior del modelo, situada en la zona inferior más al oeste aparece una franja grande de saturación de agua alta, pero como se encuentra debajo de la línea que delimita la zona que está asociada a la superficie de las escamas interpretadas por los especialistas, no se asegura con certeza su ubicación. Al igual que en el modelo anterior el comportamiento de la saturación de agua a ambos lados de la falla es distinto, esta diferencia se hace menos evidente para Veloz 3.

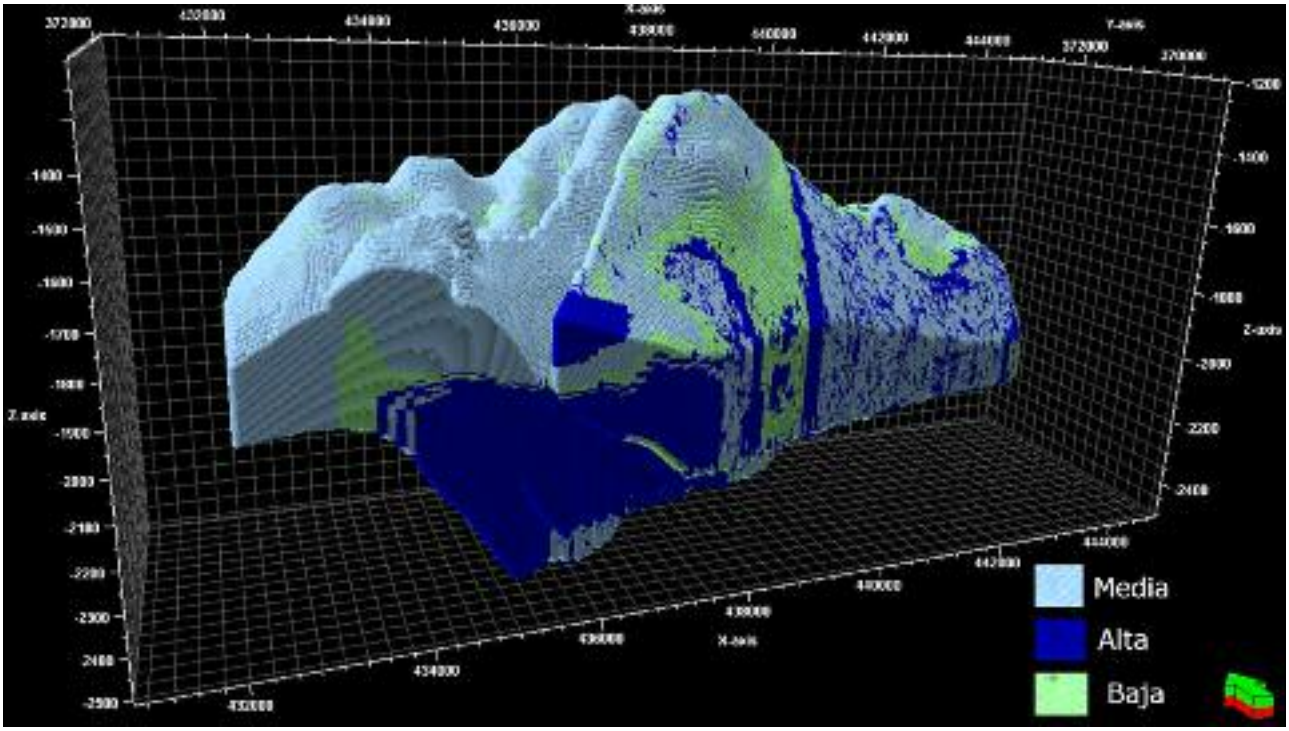

Figura 6. Modelo de saturación de agua estimado por el algoritmo *RandomForest*.

Si se analiza el modelo de calidad del reservorio mostrado en la **Figura 7** se observa cómo aparece una franja de calidad pobre en la parte central que lo divide en dos sectores, al coincidir con la información aportada por los modelos anteriores. En el sector este, las zonas con una buena calidad del reservorio comprenden casi toda la superficie de la escama al aparecer desde el tope hasta la parte inferior. En el sector situado al oeste del contacto las zonas con una buena calidad del reservorio se limitan al tope de la estructura, lo que coincide con la información aportada en los casos anteriores, los valores pertenecientes a la clase intermedia ocupan una franja grande ubicada al lado de la posible falla. Las zonas asociadas a la calidad del reservorio pobre son mínimas y aparecen casi todas situadas en la parte inferior, también en este caso aparece en la parte más al oeste una franja grande de calidad del reservorio pobre, pero al igual que en el modelo anterior como se encuentra debajo de la línea que delimita la zona que está asociada a la superficie de las escamas interpretadas por los especialistas, no asegurará con certeza su ubicación.

Para corroborar la existencia de la posible falla fue necesario recurrir al dato sísmico migrado en profundidad y a la presencia de antiguas interpretaciones, al comprobar la existencia de una zona de fallas que coincide con el contacto predicho por el modelo. También se comprobó que el rumbo de la falla predicha es NE-SW, al coincidir con el rumbo de las fallas ya interpretadas en el área.

CONCLUSIONES.

Los modelos predichos para la calidad del reservorio, la fracturación y la saturación de agua facilitaron la identificación de un comportamiento distinto en los dos sectores del modelo para las escamas Veloz 2 y Veloz 1, en el caso de la escama Veloz 3 no se identificó ninguna diferencia.

Con los análisis de la calidad alcanzada por los modelos en el entrenamiento, la seguridad mostrada en el proceso de predicción, la efectividad en la validación matemática y la coincidencia lograda con los datos geólogo-geofísicos, se determinó que los modelos más idóneos para realizar la identificación de las

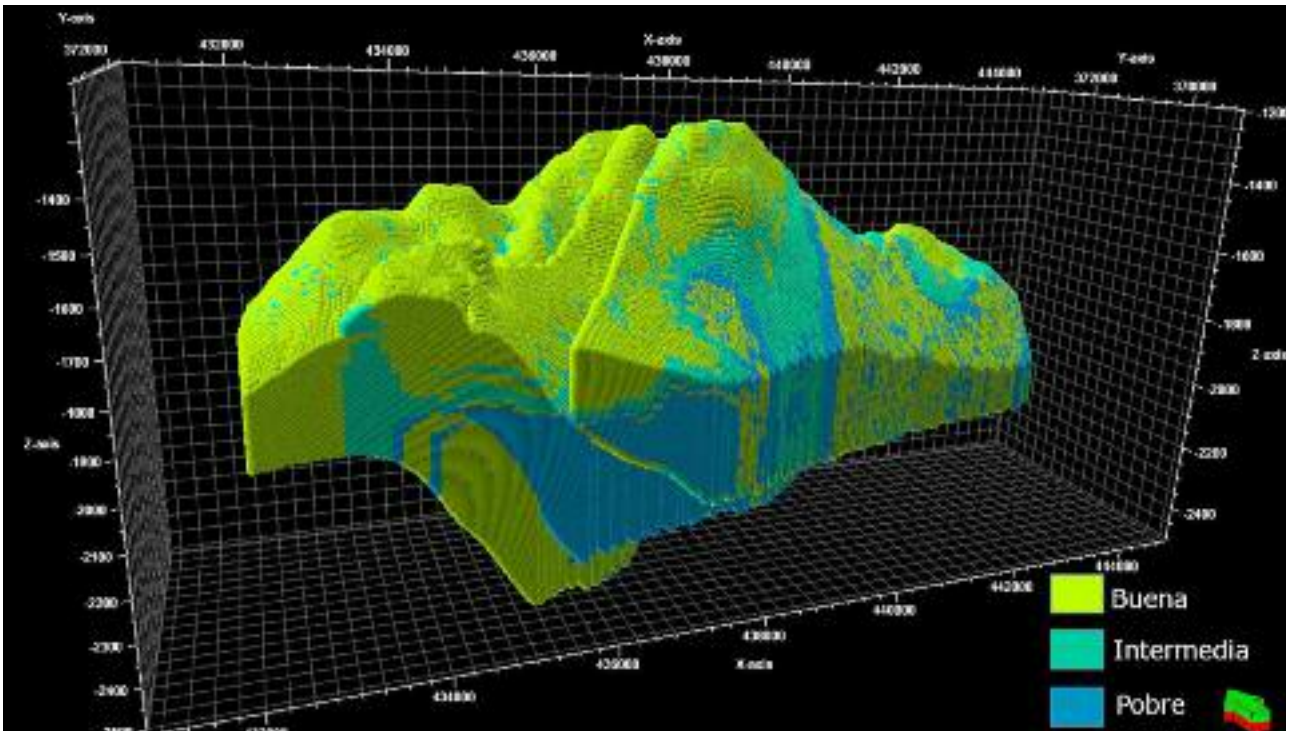

Figura 7. Modelo de Calidad del reservorio estimado por el algoritmo *RandomForest.*

áreas perspectivas son los predichos por el algoritmo *RandomForest*.

Se logró analizar la continuidad de la fracturación, la saturación de agua y la calidad del reservorio en las escamas Veloz 1, 2 y 3, lo que facilita la identificación de nuevas zonas perspectivas.

Fue identificada una falla que divide al modelo en dos sectores, con un comportamiento distinto en las escamas Veloz 2 y Veloz 1, al lograr ser validada mediante la información sísmica. La identificación de la falla es consecuente para las tres técnicas de minería de datos aplicadas.

AGRADECIMIENTOS

A todos los trabajadores de la Unidad Científico-Técnica de Base (UCTB) de Geofísica del Centro de Investigación del Petróleo de Cuba Petróleo (CUPET) y a los profesores del Departamento de Geociencias de la Universidad Tecnológica de la Habana «José Antonio Echeverría», CUJAE.

REFERENCIAS BIBLIOGRÁFICAS

Calleja Gómez, A. J., 2010, Minería de datos con WEKA para la predicción del precio de automóviles de segunda mano. Tesis en opción al grado de Ingeniero Informático (inédita), Universidad Politécnica de Valencia, Valencia, España.

Corso, C.L., 2010, Aplicación de algoritmos de clasificación supervisada usando WEKA. Universidad Tecnológica Nacional, Facultad Regional Córdoba. 11.

Castro, J.F., 2006, Fundamentos para la implementación de la red perceptron multicapa mediante software. Tesis en opción al grado de Ingeniero Informático (inédita), Univerrsidad de San Carlos de Guatemala, Guatemala.

Catalina, G.A. Introducción a las redes neuronales artificiales. 1994.

García Jiménez, M. y A. Álvarez Sierra, 2007, Análisis de datos en WEKA-Pruebas de selectividad. Universidad Carlos III. 9.

Delgado Ramos, J.R., A. Pérez Reyes, B.R. Domínguez Garcés, J.L.P. Betancourt, Y. Ta-

mayo Castellanos, M.C. Sánchez García, E.M. Gonzalez Rodríguez, R. Cruz Toledo, D. Brey Del Rey y R. Ibonet, 2019, Informe sobre la integración de datos geólogo-geofísicos del sector Yumurí-Seboruco y el sector oriental del bloque 7. Centro de Investigación del Petróleo (CEINPET), U.C.T.B. de Geofísica.

Fernández Mederos, J.C., 2016, Búsqueda directa de hidrocarburos en el yacimiento Seboruco. Tesis en opción al grado de Ingeniero Geofísico (inédita), Instituto Superior Politécnico «José Antonio Echeverría», CUJAE, La Habana, Cuba.

Zhao, T., V. Jayaram, A. Roy y K.J. Marfurt, 2015, A comparison of classification techniques for seismic facies recognition. The University of Oklahoma, ConocoPhillips School of Geology and Geophysics.

Recibido: 21 de julio de 2020
Aprobado: 7 de junio de 2021.