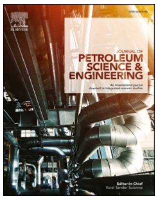
Journal of Petroleum Science and Engineering 208 (2022) 109482

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: www.elsevier.com/locate/petrol



Spatial assessment of gross vertical reservoir heterogeneity using geostatistics and GIS-based machine-learning classifiers: A case study from the Zubair Formation, Rumaila oil field, southern Iraq

Amna M. Handhal [a,b], Frank R. Ettensohn [a], Alaa M. Al-Abadi [a,b,*], Maher J. Ismail [c]

[a] Department of Earth and Environmental Sciences, University of Kentucky, Lexington, USA
[b] Department of Geology, College of Science, University of Basrah, Basrah, Iraq
[c] Basra Oil Company, Basrah, Iraq

ARTICLE INFO

Keywords:
Heterogeneity
Machine learning
Zubair formation
Rumaila oil field

ABSTRACT

The study of oil-field reservoir heterogeneity is an important task in the oil industry as it affects waterflooding, developing injection production systems, and optimizing hydrocarbon production. In this study, vertical reservoir heterogeneity was quantified using the Lorenz statistical index, empirical Bayesian kriging, and seven machine-learning classifiers (Classification and Regression Trees, Boosted Regression Trees, Random Forest, Naïve Bayes, Logistic Regression, K-Nearest Neighbors, and Support Vector Machine with three different kernels (linear, radial, and polynomial) under the geographic information system platform. The main pay zone of the Zubair Formation in the Rumaila oil field from southern Iraq was used as a case study. The degree of heterogeneity was first quantified using the Lorenz index, and a borehole-heterogeneity inventory location map was prepared according to the determined Lorenz index. Information about five factors influencing the heterogeneity, namely, porosity, permeability, volume of shale, reservoir-unit thickness, and depth to the top of reservoir unit, was collected based on available cores, nuclear magnetic resonance log, gamma-ray logs, and drilling-information logs. Factors from these sources were interpolated to show their spatial distribution using the empirical Bayesian kriging technique. The relationship between the borehole inventory map of vertical heterogeneity and the five factors was examined using the seven machine-learning classifiers. Two statistical-error measures, namely, accuracy and Cohen's kappa, were used to verify the performance of the classifiers in both training and testing stages. Results proved that Random Forest, Support Vector Machine with radial kernel function, and Logistic Regression were the best models. The probabilities of the best performance models were then interpolated and classified into five heterogeneity zones: Very low, low, moderate, high, and very high. The high-very high classes for each of these models approximately occupy 60% of the oil field and are mainly distributed in the middle and north of the field, whereas the other classes encompass about 40% of the field and mostly occur in the south. This distribution of classes is most likely related to the distribution and complexity of former depositional environments.

1. Introduction

Heterogeneity in petroleum reservoir studies is a concept used to define the variability within a particular space and/or time, and at a given scale, of a single or combination of petrophysical properties (Fitch et al., 2015). Heterogeneity is an intrinsic, pervasive, and critical property that is highly dependent on observational scales and the measurement methods used (Frazer et al., 2005). Reservoir heterogeneity occurs at different levels and scales from micrometers to hundreds of meters (Morad et al., 2010) and is commonly related to variations in depositional facies, diagenesis, and structural features (e.g., the presence of fractures and faults) (De Ros, 1998). Heterogeneity plays a major role in controlling fluid flow and recovery processes and thus has a vital influence on reservoir performance. For the implementation of a successful reservoir-development strategy, the prediction of reservoir heterogeneity is of primary importance. A quantitative

* Corresponding author. Department of Earth and Environmental Sciences, University of Kentucky, Lexington, USA
E-mail addresses: amna.handhal@uky.edu, aman.handhal@uobasrah.edu.iq (A.M. Handhal), fettens@uky.edu (F.R. Ettensohn), Alaa.Al-Abadi@uky.edu, alaa.atiaa@uobsasrah.edu.iq (A.M. Al-Abadi), mahermji@gmail.com (M.J. Ismail).

https://doi.org/10.1016/j.petrol.2021.109482
Received 23 March 2021; Received in revised form 3 September 2021; Accepted 5 September 2021
Available online 8 September 2021

assessment of heterogeneity is crucial for predicting reservoir performance during waterflooding, developing an effective injection production system, and optimizing hydrocarbon production (Handhal et al., 2020b). There are two essential types of heterogeneity: vertical and horizontal (i.e., depth and lateral variations, respectively). In the petroleum industry, geostatistical approaches are widely used to describe the two types of heterogeneity (Ahmed, 2018). For defining vertical heterogeneity, two statistical parameters are frequently used: the Lorenz coefficient L_k and the Dykstra-Parsons permeability variation V_k (Tiab and Donaldson, 2015), and geostatistical interpolation techniques (deterministic and stochastic) can be effectively used to assess the lateral heterogeneity through interpolation, as well as to extrapolate the rock and fluid characteristics of unsampled locations.

Geospatial analysis is an important GIS technique used to extract new information by integrating information from multiple, separate and disparate sources by applying a complex set of spatial operators. Geospatial analysis goes beyond simple mapping to allow research into the characteristics of places and their interrelationships. This extensive range of methods involving in geospatial analysis expands the capacity to address challenging spatial questions and lends new perspectives to decision-making. The process of integrating geospatial analysis, geostatistics, knowledge-driven, and data-driven advanced modeling techniques has opened broad prospects for studying different stochastic and deterministic petroleum-related science and engineering problems. Analysis of hydrocarbon potential and productivity (Amiri et al., 2015; Alshayef et al., 2019; Handhal et al., 2020b; Ren et al., 2020), discovery of new hydrocarbon resources (Bingham et al., 2012), and tracing paths of hydrocarbon migration (Liu et al., 2008; Rudini et al., 2018) are examples of the successful application of spatial analysis in petroleum-related fields.

In this regard, Liu et al. (2008) presented GIS-based models for searching the pathways of secondary hydrocarbon migration by considering the geologic mechanisms. The proposed algorithms were effectively implemented in the modeling of secondary pathways in the northern Songliao Basin, northeast China. The findings of modeling agreed well with the drilling data and demonstrated the resilience of the approaches suggested. Bingham et al. (2012) proposed a GIS-based multicriteria (MCE) method for petroleum exploration based on fuzzy logic to produce a favorability map of potential exploration areas and a case study from northern south America was taken to show the potential new exploration areas in the Cretaceous-Paleogene and Miocene-Holocene. They concluded that it is possible to use the suggested GIS approach in an exploratory scenario and to other locations of the world. Amiri et al. (2015) mapped the hydrocarbon resource potential using GIS-based two statistical models namely, frequency ratio and evidential belief functions. A case study in the Red River petroleum system in the Canadian Williston Basin in southeastern Saskatchewan of Canada is selected to assess the feasibility of the proposed modeling techniques. Model results are evaluated by success rate and prediction rate efficiency curves. The resulting hydrocarbon potential map has led to the delineation of high-potential areas which cover approximately 15% of the study area. Rudini et al. (2018) modeled the migration of hydrocarbon using GIS and used the top of Group E horizon of the northeast Malay basin as an illustrated example. They utilized the seismic data, well log data, lithology, and simple overlay technique to map the hydrocarbon migration and showed that the produced map is well-matched with previous studies which use advanced technology to carry out the analysis. Alshayef et al. (2019) have tried to use the geophysical, geological, and remote sensing data incorporated into ArcGIS software to delineate promising zones of hydrocarbon potential in the Masila oil field, Yamen. They used lineaments as the main theme to produce a map of hydrocarbon potentiality beside the seismic, gravity, magnetic, and geological map of the considered basin. The resulting potential map of their study was classified into four zones: low, moderate, high, and very high and verified with oil fields and existing productive wells which showed a positive correlation. Their main conclusion confirmed that spatial models are significant for hydrocarbon potential resource planning and management. Handhal et al. (2019) used three machine learning models, specifically, support vector machine, naïve Bayes, and random forest with GIS to delineate the tar mat occurrence in the upper part of Zubari Formation at Rumila oil field, Iraq. Applying the models suggested that the random forest was the best performance model followed by a support vector machine with a polynomial kernel. Their findings confirmed that GIS-based machine learning models offer an easy and costly way of avoiding the drilling wells where the tar is expected to occur. Ren et al. (2020) applied a Bayesian network algorithm to predict the spatial distribution of oil and gas resources and used the first member of Dongying Formation of Nanpu Depression Bohai Bay Basin, China as a case study. They utilized 222 exploratory wells, basin simulation, seismic interpretation, and other auxiliary data to train the Tree Augmented Bayesian Network structure and mapped the hydrocarbon-bearing posterior probability of the reservoir member. Results of this study showed that the Bayesian network captures essential spatial characteristics of hydrocarbon accumulations and accurately predicts the spatial distribution of oil and gas resources, which can help to manage the reservoir units and initiate successful drilling programs. Handhal et al. (2020b) developed a GIS-based hybridization of Shannon's entropy and the technique for order preference by similarity to an ideal solution (TOPSIS) model to map the hydrocarbon productivity of the middle reservoir unit of the Nahr Umr Formation in the Luhais oil field in southern Iraq. They quantified the heterogeneity of the reservoir unit using the Lorenz coefficient and Dykstra-Parsons permeability indices and used the hydraulic flow unit concept to overcome the heterogeneity problem in the spatial model formulation. They utilized the ordinary kriging technique to interpolate the seven selected petrophysical properties (porosity, unit thickness, volume of shale, bulk volume of water, total water saturation, hydrocarbon saturation, and bulk volume of hydrocarbon) and entropy information theory to assign weights of these properties to use in the TOPSIS model to demarcate the hydrocarbon productivity across the reservoir unit. The major conclusion of their study revealed that the spatial model offers a simple approach to map hydrocarbon productivity that can be successfully used by reservoir management, geologist, and reservoir engineers for drilling new productive boreholes with the least effort and expense.

From a review of previous studies, it can be said that there is no study so far to study the spatial distribution of vertical heterogeneity using GIS and machine learning-based techniques. Therefore, in this study, the spatial distribution of vertical heterogeneity of the DJ unit of the main pay zone of the Zubair Formation (Fig. 1), southern Iraq, were modeled using the statistical Lorenz coefficient (L_k), the Empirical Bayesian kriging interpolation (EBK) technique, and advanced master machine-learning algorithms. The objective was to map the spatial distribution of vertical gross heterogeneity across the field to better manage the reservoir unit by modeling the relationship between the L_k as the target variable and five reservoir-related properties, namely, the depth to reservoir unit top, the average of unit thickness, and the average of porosity, permeability, and volume of shale as predictors.

2. Study area location and geological setting

Rumaila is a supergiant oil field found in 1953 by Basrah Petroleum Company, situated 50 km west of the city of Basrah and 30 km west of Zubair oil field in southern Iraq (Fig. 2). It covers an area of 1600 km^2 and represents an 80-km-long, north-south anticline, extending from the Iraqi-Kuwait border in the south into the West Qurna oilfield to the north. The topography of the field is almost smooth, sloping gently from about 70 m above mean sea level in the south to near sea level in the north. The field is composed of two domes (north and south). It is a gentle sloping longitudinal anticline and is stretching around 83 km long and 12 km wide. In June 1959, the north Rumaila structure was drilled as a northern step-out of the Rumaila axis to delineate the northern

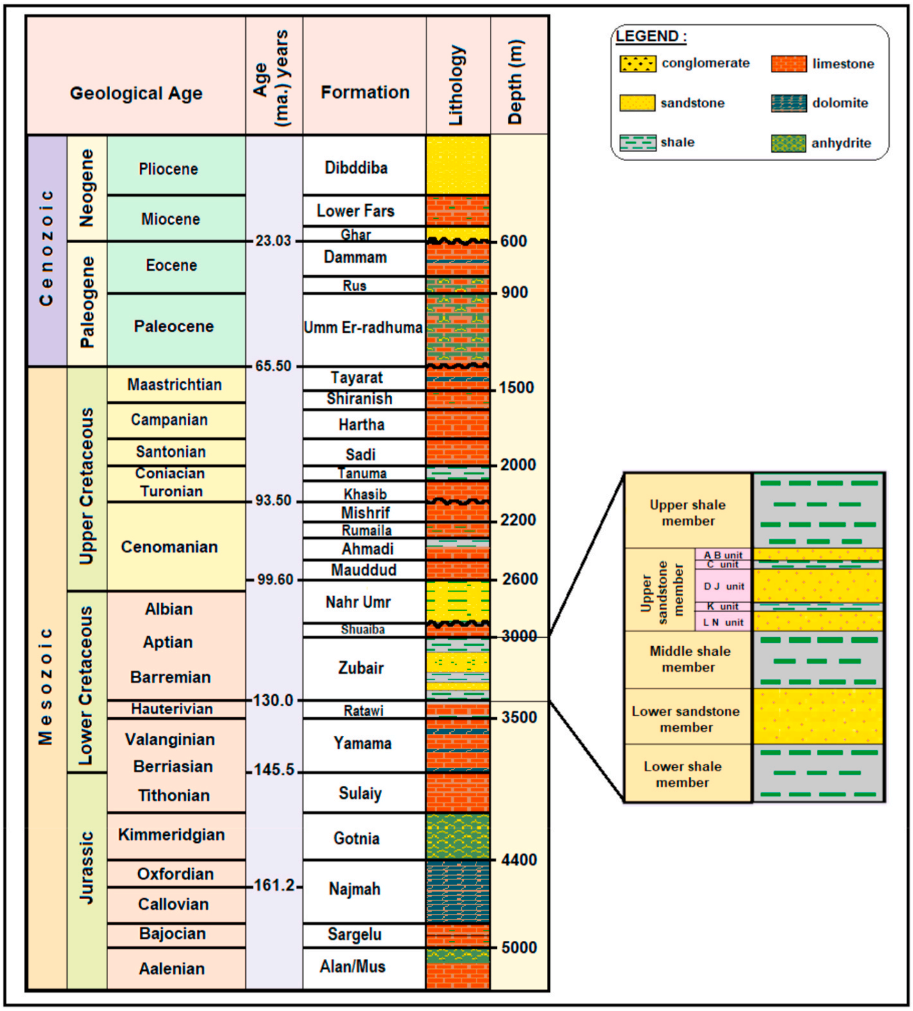


Fig. 1. Stratigraphic column of North Rumaila oil field (after Handhal et al., 2019, and Elsevier, License No. 4970940455936, Dec. 16, 2020).

plunge and explore the depth of carbonate prospects in the Upper Cretaceous Mishrif Formation, as well as for investigating the lower sandstone member of Zubair Formation below the main pay zone. The North Rumaila structure is approximately 42 to 11 km in width and gently slopes southwards to form a saddle that separates it from the south Rumaila structure (Al-Ansari, 1993) (Fig. 2). From a geological point of view, the Rumaila oilfield is located in the Mesopotamian Plain, an area of subsiding, Quaternary, terrestrial, floodplain sedimentation that sits atop the Mesopotamian foredeep (Fouad, 2010). At one time, the present foredeep area was part of a stable shelf area on the north-eastern margin of the Arabian plate, and from Permian to Late Cretaceous time, the area was a passive-margin, epicontinental basin, which experienced periods of rifting and subsidence, related to the opening of the Neo-Tethys Sea (Jassim and Goff, 2006). By Late Cretaceous time, closure of the Neo-Tethys had begun with major thrusting of ophiolites and deep-water sediments against the western margin of the Arabian plate, at which time the Mesopotamian epicontinental basin became a distal part of the Zagros foreland basin (Fouad, 2010; Fouad and Sissakian, 2011; Jassim and Goff, 2006). Closure of the Neo-Tethys continues to this day with the collision of the Arabian and Eurasian (Iranian) plates with the Mesopotamian and Persian Gulf basins representing terrestrial and marine remnants, respectively, of the Zagros foreland basin (Fouad and Sissakian, 2011). Hence, the Mesopotamian foredeep is an unstable, actively subsiding area with mainly subsurface structures, including folds, faults, and diapiric structures. Two of these faults, the northeast-southwest, Takadid-Qurna and Al-Batin faults are reactivated Precambrian transverse faults that define the Zubair fault block or subzone of the Mesopotamian basin (Jassim and Goff, 2006) (Fig. 2), in which the Rumaila oilfield is situated. The Rumaila oilfield and others in the area, formed over a series of north-south-oriented, Infracambrian rift basins containing salt, which were remobilized by east-west compression during Tertiary collision of the Arabian and Eurasian (Iranian) plates (Jassim and Goff, 2006; Fouad, 2010; Fouad and Sissakian, 2011).

The stratigraphic column from the Rumaila field represents sedimentary rocks, ranging in age from Late Jurassic to Recent (Jaffar, 2018) and mainly comprises of cycles of clastic, carbonate and evaporitic rocks (Fig. 1). The most significant hydrocarbon system in this stratigraphic column is the Early Cretaceous–Miocene petroleum system. In this petroleum system, the Sulaiy and Yamama formations represent source rocks; the Tanuma, Shranish, and Rus formations form the sealing rocks; and the Yamama, Zubair, Nahr Umr, and Mishrif formations represent the reservoir rocks (Fig. 1). (Aqrawi et al., 2010). The Zubair Formation contains the most significant Lower Cretaceous cycle in Iraq, and it is primarily composed of fluvio-deltaic and marine sandstones (Fig. 1), which are Hauterivian-early Aptian in age (Bellen et al., 1959). The average thickness of the formation is 425 m, and the contact of the formation with adjacent formations is mostly gradational. It is overlain by Shuaiba Formation (limestone and dolomite) and is underlain by Ratawi Formation (interbedded limestones and shales). In type locality from the Zubair subzone, the formation is divided into five units: the upper shale member, the upper sandstone member (the main reservoir or main pay), the middle shale member, the lower sandstone member,

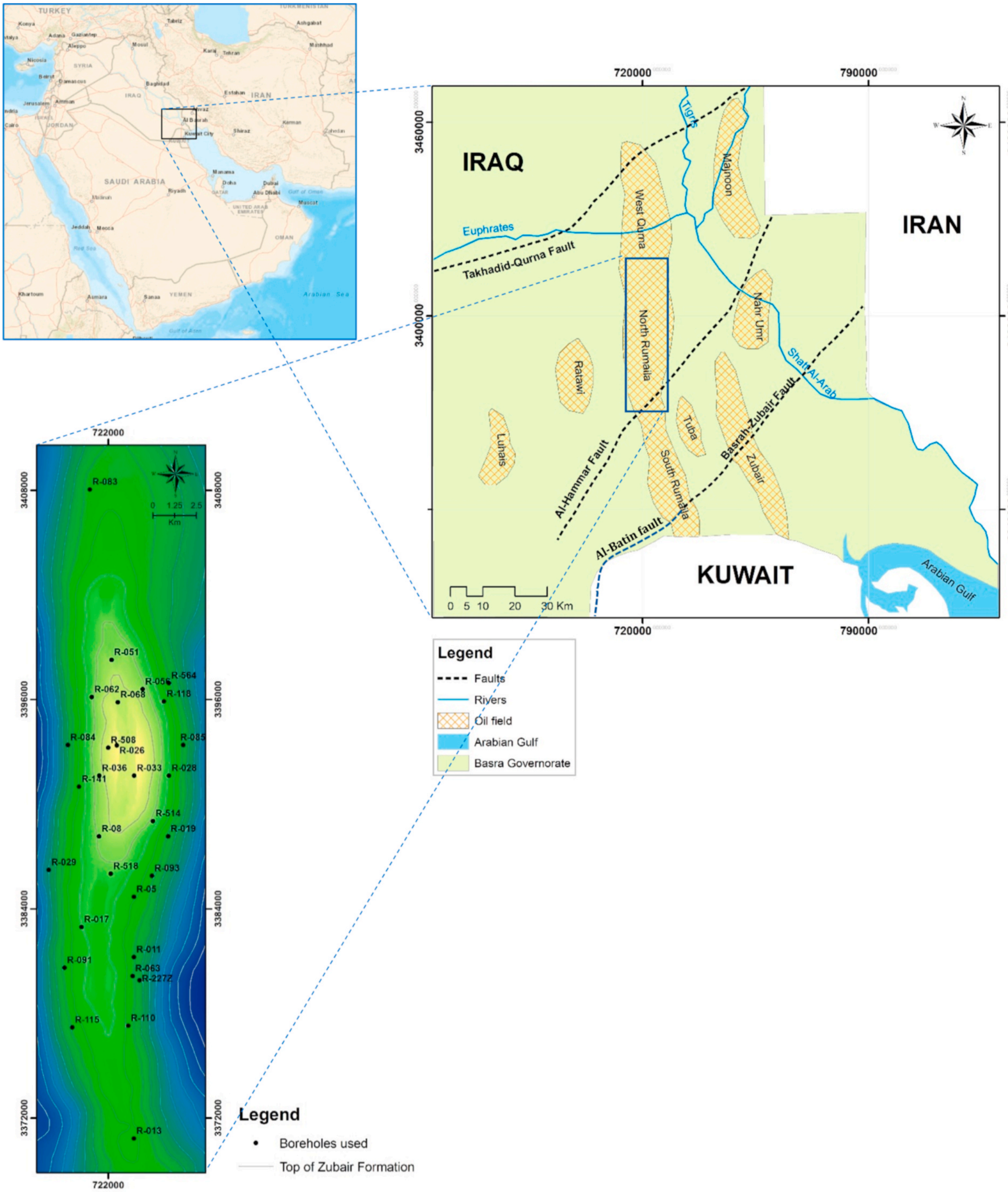


Fig. 2. Study area relative to other oilfields and transverse fault zones in the Zubair subzone of the Mesopotamian foredeep basin.

and the lower shale member (Owen and Nasr, 1958). The present study focuses on the upper sandstone member of the Zubair Formation, which is located at an average depth of approximately 3150 m below mean sea level. It consists of mainly sandstones with some interbedded shales. The total thickness of the reservoir is about 145 m, and it contains three reservoir units, namely AB, DJ, and LN from top to bottom, separated by two isolating units C and K (Figs. 1 and 3). The smallest and most widespread reservoir unit is the AB Unit, whose thickness ranges from 2.8 to 14 m and mainly consists of sandstone with thin layers or lenses of silt or shale. The percentage of sand typically increases toward western parts of the field. Unit C is an isolating unit between sandstone units AB and DJ and mainly consists of shale and siltstone; its thickness ranges from 3 to 8 m and increases toward eastern and northern parts of the field. The thickest reservoir unit, the DJ unit, is about 46–66 m thick and consists of sandstone interspersed between two layers of siltstone. Because of its very good reservoir characteristics, it represents the main part of the reservoir in the upper sandstone member (the main pay). In contrast, the K unit represents a barrier between DJ and LN units, and the lithology of the unit changes from siltstone at the top and along the western margin of the structure to the shale and silty shale along the northern and eastern margins of the field; its thickness ranges from 3 to 9 m. The K unit decreases in thickness in the southern, western, and southwestern parts of the field. The thickness of the LN unit ranges from 30 to 53 m and represents sandstone with interbedded layers and lenses

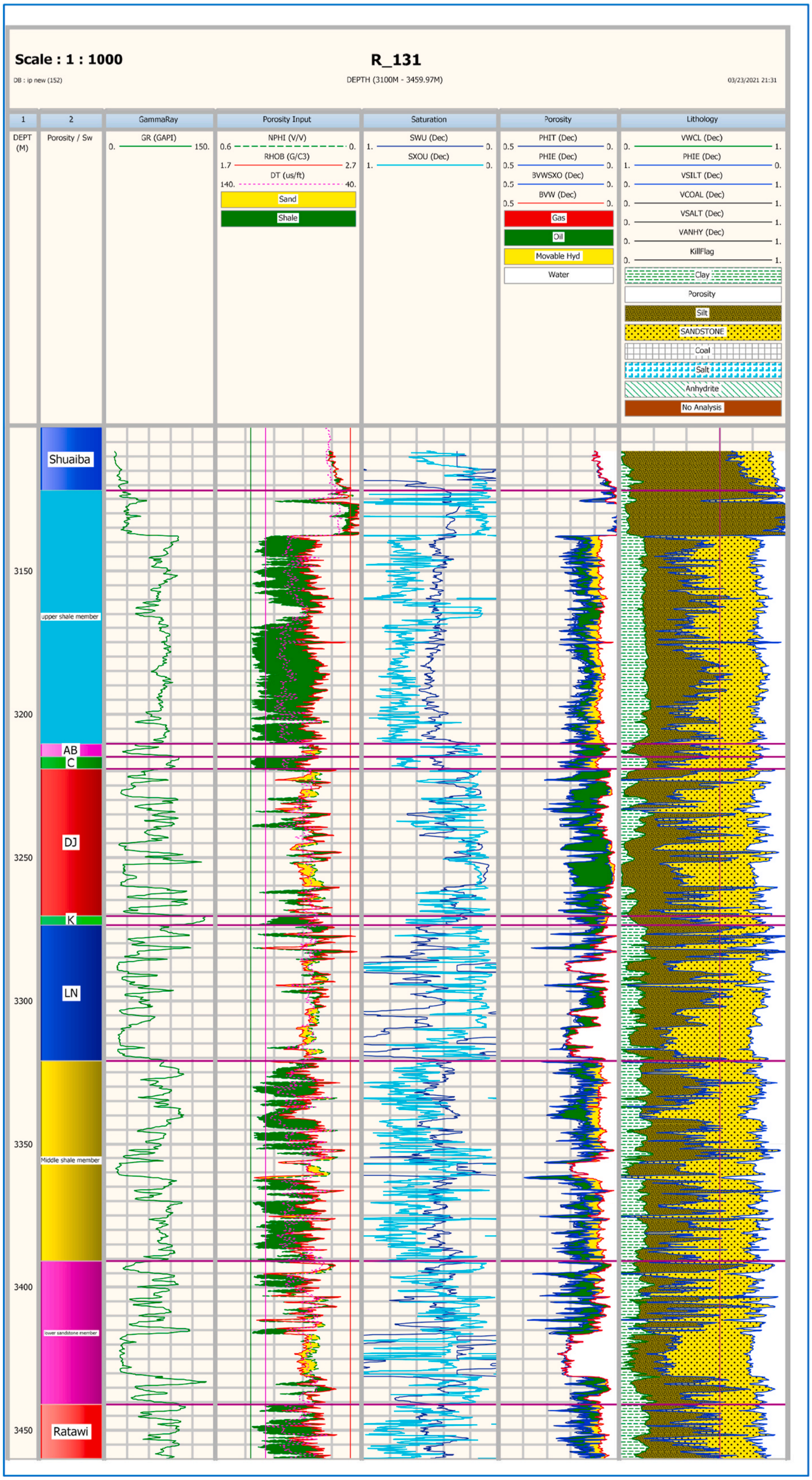


Fig. 3. The main members of Zubair Formation in the North Rumaila oil field, along with their gamma-ray, porosity-input, saturation, porosity and lithologic log characteristics.

of shaly siltstone and shale, which are greater in number than those found in the DJ unit. The siltstone and shale content of the LN unit often increases towards the base of the unit, reflecting its proximity to the underlying middle shale member. The percentage of sand in the LN unit increases toward the western margin of the structure.

3. Methodology

In mapping the gross spatial vertical heterogeneity of the DJ reservoir unit, six steps were followed (Fig. 4): (i) Collecting available core porosity and permeability data from 26 boreholes in the DJ unit, nuclear magnetic resonance (NMR) data from five boreholes, the well-drilling information from 58 boreholes, and gamma-ray logs from 58 boreholes; (ii) using porosity and permeability data, the degree of heterogeneity was calculated using the Lorenz coefficient L_k; (iii) preparing borehole-heterogeneity inventory map based on the calculated $\mathrm{L_k}$ by assigning 0 (homogenous) for boreholes having $L_k < 0.5$ and 1 (heterogeneous) for boreholes having $L_k > 0.5$; the total number of assigned boreholes was then divided into two sets after rebalancing classes, because the number of 0-coded points was much lower than 1-coded points with a 0.2/0.8 ratio: 70% for training and 30% of data for testing the models used; (iv) based on the available data, five factors were selected as influencing factors affecting the heterogeneity of the reservoir units, namely, the average porosity, permeability, volume of shale, unit thickness, and depth to the top of the reservoir unit; these five factors were initially interpolated using the Bayesian kriging technique to reveal their spatial distribution throughout the study area; (v) the relationships between borehole location (as the dependent variable; response) and the factors influencing heterogeneity were modeled using seven machine-learning algorithms, namely, (Classification and Regression Trees [CART], Boosted Regression Trees [BRT], and Random Forest [RF], Naïve Bayes [NB], Logistic Regression [LR], K-Nearest Neighbors [KNN], and Support Vector Machine [SVM] with three different kernels (linear, radial, and polynomial); two error-measuring statistics, specifically, the accuracy and Cohen's kappa statistics were used to assess the models performance in both training and testing stages, and (vii) using the three best performance models, the probability of heterogeneity were estimated and mapped for the unit reservoir. A detailed description of these steps is outlined below.

4. Material and methods

4.1. Data used

The φ (fraction) and k (md) core data from the 26 boreholes were used in this study to calculate the degree of heterogeneity of the reservoir units using the L_K method. The total number of core data used in this study was 1553. A statistical summary of the core data used is presented in Table 1, and the histograms of these parameters are shown in Fig. 5. Besides core data, there are five boreholes, namely R-227z, R-514, R-564, R-508, and R-518, which have only NMR-log data and conventional-log data. The k values for these boreholes were calculated using the Schlumberger-Doll-Research (SDR) model according to the following equation (Yarmohammadi et al., 2020):

$$k = C_1 \times \varphi^{m1} \times T_{2lm}^{n1} \tag{1}$$

where φ is the total porosity (%), T_{2lm} (ms) is the logarithmic mean of NMR T_2 spectra, and m, n, and C are statistical parameters from the model.

To investigate the relationship between the log(k), φ, and $\mathrm{V_{sh}}$,

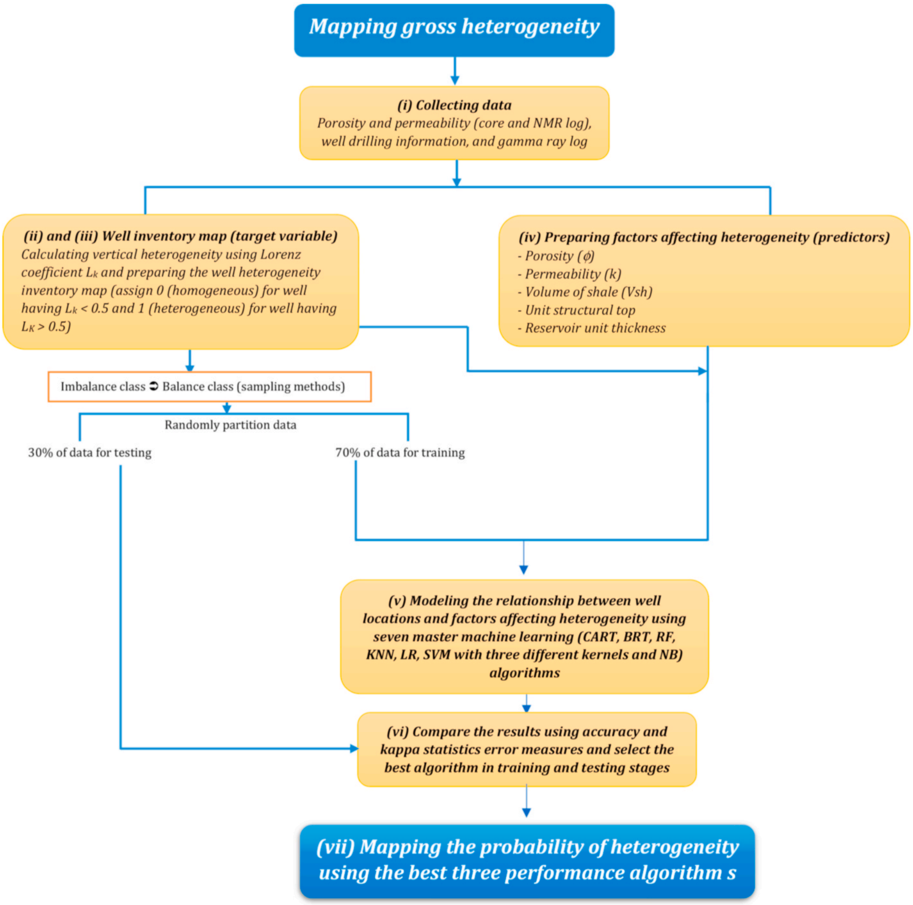

Mapping gross heterogeneity
(i) Collecting data
Porosity and permeability (core and NMR log), well drilling information, and gamma ray log
(ii) and (iii) Well inventory map (target variable)
Calculating vertical heterogeneity using Lorenz coefficient Lk and preparing the well heterogeneity inventory map (assign 0 (homogeneous) for well having Lk < 0.5 and 1 (heterogeneous) for well having LK > 0.5)
(iv) Preparing factors affecting heterogeneity (predictors)
- Porosity (ϕ)
- Permeability (k)
- Volume of shale (Vsh)
- Unit structural top
- Reservoir unit thickness
Imbalance class ➲ Balance class (sampling methods)
Randomly partition data
30% of data for testing
70% of data for training
(v) Modeling the relationship between well locations and factors affecting heterogeneity using seven master machine learning (CART, BRT, RF, KNN, LR, SVM with three different kernels and NB) algorithms
(vi) Compare the results using accuracy and kappa statistics error measures and select the best algorithm in training and testing stages
(vii) Mapping the probability of heterogeneity using the best three performance algorithm s

Fig. 4. Steps adapted in this study for mapping gross heterogeneity of the DJ reservoir unit.

Table 1
Statistical summary of φ (fraction) and k (md) for the available core and NMR log data.

Parameter	N	Min.	Max.	Mean	St. Dev.	C_v (%)	Skewness
φ (core)	1553	0.001	0.286	0.174	0.056	32.33	−0.90
φ (NMR)	1577	0.008	0.298	0.141	0.014	29.53	0.20
k (core)	1553	0.000	6405	384.4	644.4	167.6	3.75
k (NMR)	1577	0.000	5568	280.3	731.6	261.05	4.06

N: number of measurements; St. Dev.: standard deviation; C_v: coefficient of variation.

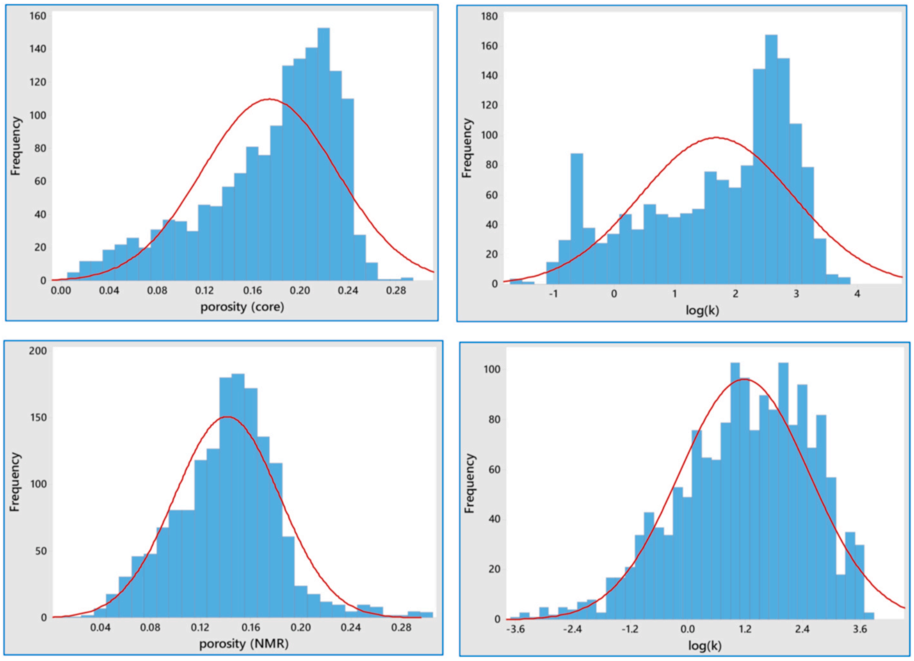


Fig. 5. Histograms of porosity and log of permeability from core- (upper) and NMR-derived logs (lower).

correlation and regression analyses of three wells, namely, R-083, R-508, and R-013 that distribute at the northern, middle, and southern parts of the oil field, was implemented (Table 2).

Table 2
Regression and correlation results.

Well No.	Log(k) vs. φ	Regression Equation	Log(k) vs. φ and V_{sh}	Regression Equation
R-013	$R^2 = 0.685$ (strong)	Log(k) = −1.271 + 17.95φ	$R^2 = 0.741$ (strong)	Log(k) = -0.427 + 14.47φ-5.79 V_{sh}
R-508	$R^2 = 0.446$ (weak)	Log(k) = -1.451 + 19.54φ	$R^2 = 0.605$ (moderate)	Log(k) = = 0.470 + 14.41φ-4.24V_{sh}
R-083	$R^2 = 0.618$ (moderate)	Log(k) = -1.901 + 15.53φ	$R^2 = 0.626$ (moderate)	Log(k) = = -1.993 + 15.39φ+1.70V_{sh}

Correlation table

Well No.	Correlation Coefficient (r)		
R-013		Log(k)	φ
	φ	0.828	
	V_{sh}	−0.660	−0.560
R-508		Log(k)	φ
	φ	0.668	
	V_{sh}	−0.634	−0.403
R-083		Log(k)	φ
	φ	0.786	
	V_{sh}	0.152	0.076

4.2. *Degree of heterogeneity calculation and borehole heterogeneity inventory map*

L_k is a statistical measure of heterogeneity calculated by plotting the cumulative flow capacity (a product of the average permeability and reservoir unit thickness) on the y-axis against the cumulative storage capacity (a product of average porosity and thickness for the same reservoir unit) on the x-axis (Handhal et al., 2020b). The L_k value varies from 0 to 1, and the reservoir will have a uniform distribution of permeability (entirely homogeneous) if $L_k = 0$, but if $L_k = 1$, the reservoir is deemed to be entirely heterogeneous (Tiab and Donaldson, 2015). After calculating L_k from the available data (from core and NMR logs) (see Table 3 and Fig. 6), the boreholes were classified into two groups: those having $L_k < 0.5$ and those having $L_k > 0.5$. Boreholes that have $L_k < 0.5$, were assigned a 0 code (homogeneous), whereas boreholes that have $L_k > 0.5$ were assigned 1 code (heterogeneous). These codes were then used in the classification problem solved in this study to reveal the probability of spatial heterogeneity in the studied reservoir unit. The number of boreholes with a 0 code is much lower than the number of boreholes with a 1 code (the proportion is 0.2/0.8) (Table 3). In machine learning, this difference is referred to as class imbalance, and such a class imbalance has been found to have a major adverse effect on training machine-learning classifiers (Japkowicz and Stephen, 2002). An imbalance like this influences both convergences during the training phase and the generalization of a model on a test set (Buda et al., 2018). Methods to cope with this imbalance for master machine-learning classifiers are well established (Chawla, 2009; Mazurowski et al., 2008). The most consistent and popular techniques are the use of sampling methods and these methods operate on the data itself (rather than

Table 3
L_k coefficients for the used boreholes.

Borehole	k (md)	φ	L_k	Status	Borehole	k (md)	φ	L_k	Status
R-05	254.2	0.138	0.580	1	R-078	98.1	0.162	0.768	1
R-08	510.4	0.197	0.400	0	R-083	52.0	0.160	0.735	1
R-011	191.1	0.168	0.344	0	R-084	501.5	0.159	0.720	1
R-013	510.3	0.199	0.440	0	R-085	146.2	0.158	0.590	1
R-017	422.5	0.187	0.550	1	R-091	477.5	0.197	0.510	1
R-019	205.1	0.139	0.570	1	R-093	252.7	0.159	0.570	1
R-026	148.0	0.171	0.570	1	R-110	667.1	0.177	0.460	0
R-028	197.7	0.158	0.740	1	R-115	447.2	0.187	0.340	0
R-029	311.4	0.174	0.510	1	R-118	984.9	0.169	0.560	1
R-033	190.6	0.179	0.540	1	R-141	23.1	0.157	0.760	1
R-036	762.6	0.182	0.600	1	R-227Z	241.5	0.145	0.590	1
R-051	294.4	0.184	0.640	1	R-508	29.8	0.109	0.640	1
R-056	324.9	0.175	0.590	1	R-514	74.1	0.130	0.580	1
R-062	589.5	0.193	0.555	1	R-518	855.4	0.150	0.570	1
R-063	784.8	0.197	0.410	0	R-564	45.3	0.208	0.700	1
R-068	193.8	0.193	0.605	1					

0 means homogeneous; 1 means heterogeneous.

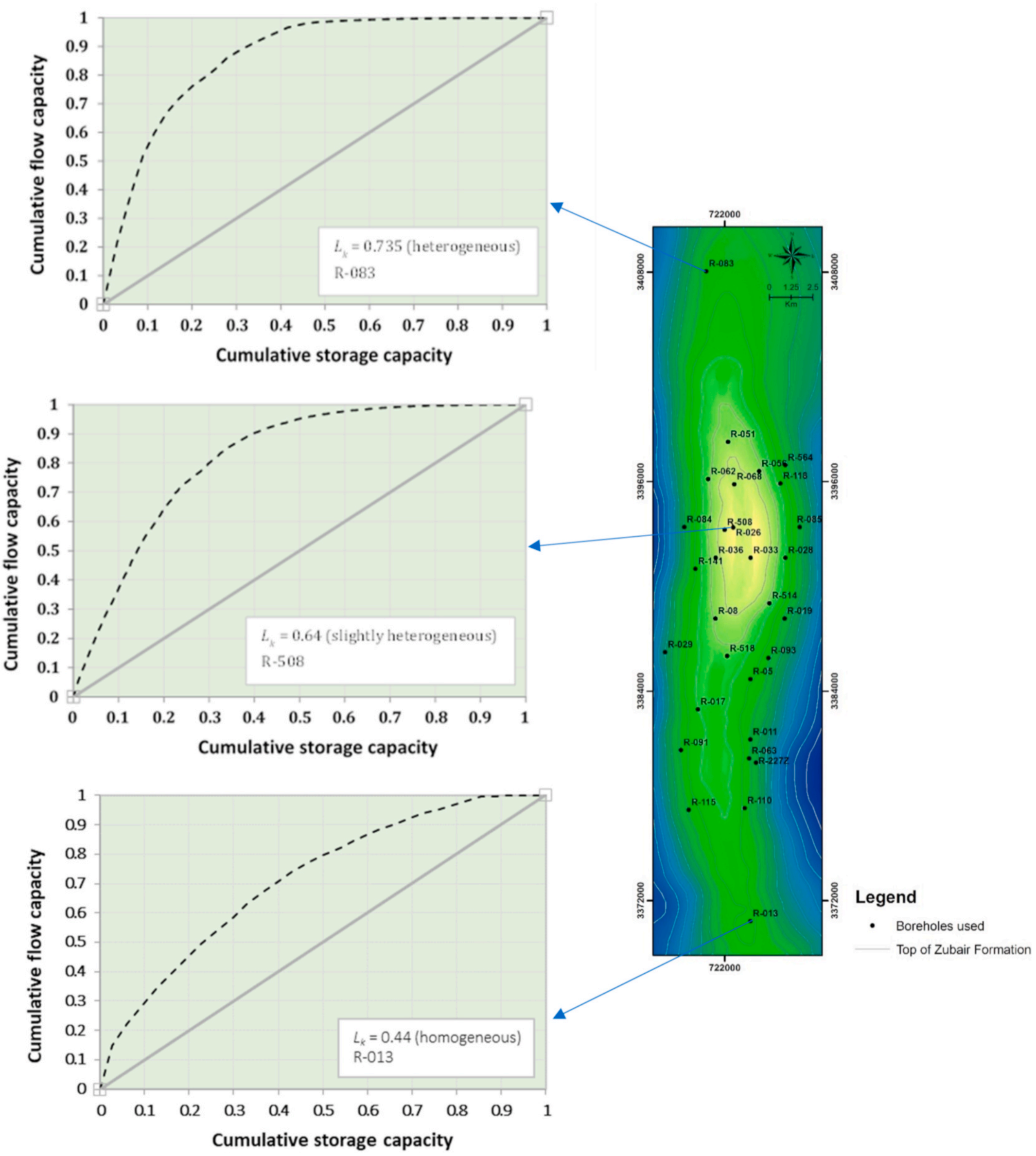


Fig. 6. Lorenz coefficient L_k for selected boreholes across the oil field.

the model) to increase its balance. In general, these methods are structured to use a process that adjusts an imbalanced dataset to a balanced one. This adjustment happens by modifying the scale of the original dataset to insure the same balancing ratio (Al-Abadi and Alsamaani, 2020; Handhal et al., 2019). The options available for adjustment are oversampling, undersampling, both (over and under), synthetic data generation, and cost-sensitive learning. The class balance is also answered at the level of the classifier itself. In such instances, the algorithms are updated by, for example, introducing various weights to misclassify examples from different classes (Zhou and Liu, 2005) or directly changing the probabilities of the prior class (Lawrence et al., 1998). In this study, the "both" scheme was used so that the minority and majority classes are oversampled (with replacement) and undersampled (without replacement), respectively. For the class-unbalance issue in this study, we used the ROSE package in R statistical software. By trial-and-error procedure, the number of examples was optimized to 60 (30 for boreholes having a homogeneous code and 30 for boreholes with a heterogeneous code). Finally, the balanced dataset was randomly partitioned into two sets: training and testing. Seventy percent (70%) of the data was used for training the seven machine-learning models, whereas the remaining (30%) was used for validating the models.

4.3. Factors affecting heterogeneity

In this study and depending on data available in the first place and the nature of the problem-solving, five factors were taken into consideration to model the spatial distribution of heterogeneity: the average of φ, k, and volume of shale (V_{sh}), in addition to depth to the top of reservoir unit and reservoir unit thickness. It is well known that reservoir heterogeneity, a measure of spatial porosity/permeability variation, is a lithology property, resulting from a combination of sedimentary and diagenetic processes. To reveal the spatial distribution of φ and k, EBK interpolation techniques were used in this study. EBK is a special kind of stochastic kriging technique, which utilizes a linear-estimation procedure to estimate a value at unsampled locations. In principle, the technique assumes that the value at the unsampled location is estimated by (Kelkar et al., 2002):

$$X^*(\overrightarrow{u_o}) = \sum_{i=1}^{n} \lambda_i X(\overrightarrow{u_i}) \tag{2}$$

where $X^*(\overrightarrow{u_o})$ and $X(\overrightarrow{u_i})$are the estimated value at the unsampled location and the value at the neighboring location, respectively; $\overrightarrow{u_i}$ and λ_i is the weight assigned to the neighboring value.

In simple words, kriging uses the spatial-variance model to estimate the variable in locations where there are no values for this variable, based on the measured values at neighboring locations, and it takes into account the covariance function between the known and unknown data (Al-Mudhafar, 2019). The purpose of the estimation procedure is to estimate the weights allocated to the individual points in the neighborhood. The spatial relationship between locations without samples and the neighboring sample values, as well as the relationship among the neighboring values, depends on these weights. These relationships are obtained from the modeling of the variogram (Rivoirard, 2005). The downside to kriging is the neglecting of uncertainty in variogram parameters (sill, range, and nugget) when building the covariance function. Therefore, EBK has been proposed to account for parameter uncertainty among the variogram parameters (Al-Mudhafar, 2019; Al-Mudhafar and Hakim, 2015). In EBK, the variogram parameters are calculated automatically by a process of subsetting and simulation (Krivoruchko, 2011). The semivariogram parameters in EBK are calculated using restricted maximum likelihood, unlike other kriging methods which use weighted least squares. The input data is first divided into overlapping subsets of a given size (the default is 100 points per subset in the Geostatistical extension of ArcGIS software used here to interpolate process), Semivariograms are calculated in each subset in the following way (Gribov and Krivoruchko, 2020): (i) The data in the subset is used to build a semivariogram. (ii) New data is unconditionally simulated at each of the input positions in the subset using this semivariogram as a model. (iii) The simulated data is used to create a new semivariogram. (iv) The steps 2 and 3 are repeated for a specified number of times. The semivariogram calculated in phase 1 is used to simulate a new set of data at the input locations in each repetition, and the simulated data is then used to estimate a new semivariogram. For a given distance h, EBK of ArcMap GIS Geostatistical extension supports three semivariograms: Power ($\gamma(h) = Nugget + b|h|^{\alpha}$, Linear $\gamma(h) = Nugget + b|h|$, and Thin Plate Spline $\gamma(h) = Nugget + b\left|h^2\right|{}^*ln(|h|)$. The *nugget* and slope (b) must be positive and the power α must be between 0.25 and 1.75. These three models do not have a rang or sill parameter because the functions have not upper bound (Krivoruchko, 2011). EBK also offers different normal transformation to modify the skewed normal properties with the choice of two base distribution: Empirical and log Empirical. Instead of using an inherent random function, a simple kriging model is used when the transformation is implemented. The parameter distribution shift to Nugget, Sill, and Range as a result of these modification (Krivoruchko, 2011). The math behind EBK can be found elsewhere (e.g., Diggle and Lophaven, 2006; Nowak et al., 2010).

The interpolated surfaces of φ and k using EBK are shown in Fig. 7a and b. The Geostatistics package from ESRI ArcGIS 10.7 was used to interpolate these petrophysical properties in the study area. The semivariogram analyses for the used factors were shown in Table 4. We used the root mean squared error (RMSE) as an error statistic to select the best performance model. It is mathematically defined as (Li et al., 2020):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{3}$$

where y_i and x_i are the simulated and observed values, respectively. The smaller the value of this statistic, the better performance of the model is.

The other factors used in the analysis include V_{sh}, the depth to the top of reservoir unit, and unit thickness. These variables were investigated to establish a statistical proxy relationship which can be used for prediction of reservoir heterogeneity where direct observation of heterogeneity is not possible. The V_{sh} was estimated using data from gamma-ray logs available from the 58 boreholes in the study area. Data for reservoir unit thickness were gathered from drilling-information logs for the same boreholes used to calculate V_{sh}. The interpolated surfaces for V_{sh} and unit thickness were also generated by EBK using the Geostatistics tool in ArcGIS 10.7 software (Fig. 7c and d). The depth to the top of the reservoir unit was digitized from a hard copy of this parameter from other work (e.g., Almalikee and Al-Najm, 2019; Handhal et al., 2020b; Jaffar and Abdulnaby, 2018), Fig. 7e.

For the φ and k (the major factors affecting heterogeneity), the arithmetic and geometric means of factors were used to quantify the effect of averaging on the interpolated surfaces (Table 3) while the arithmetic average of V_{sh} and reservoir unit factors were used to produce the interpolated surfaces of these factors.

4.4. Machine-learning classifiers

4.4.1. CART

CART is a term invented by Leo Breiman (Breiman et al., 1984) to refer to decision-tree (DT) algorithms that can be used for classification and regression modeling problems. It is a binary, recursive, partitioning procedure capable of processing continuous and nominal attributes as targets and predictors (Wu and Kumar, 2009). CART constructs a predictive model by splitting the data in the root node into two "children," and each of the children is split into "grandchildren" in turn. Trees are grown to full size without use of a stopping rule; when no more splits are possible due to lack of data, the tree-growing process ceases (Al-Abadi et al., 2019; Wu and Kumar, 2009). The full tree is then pruned back to

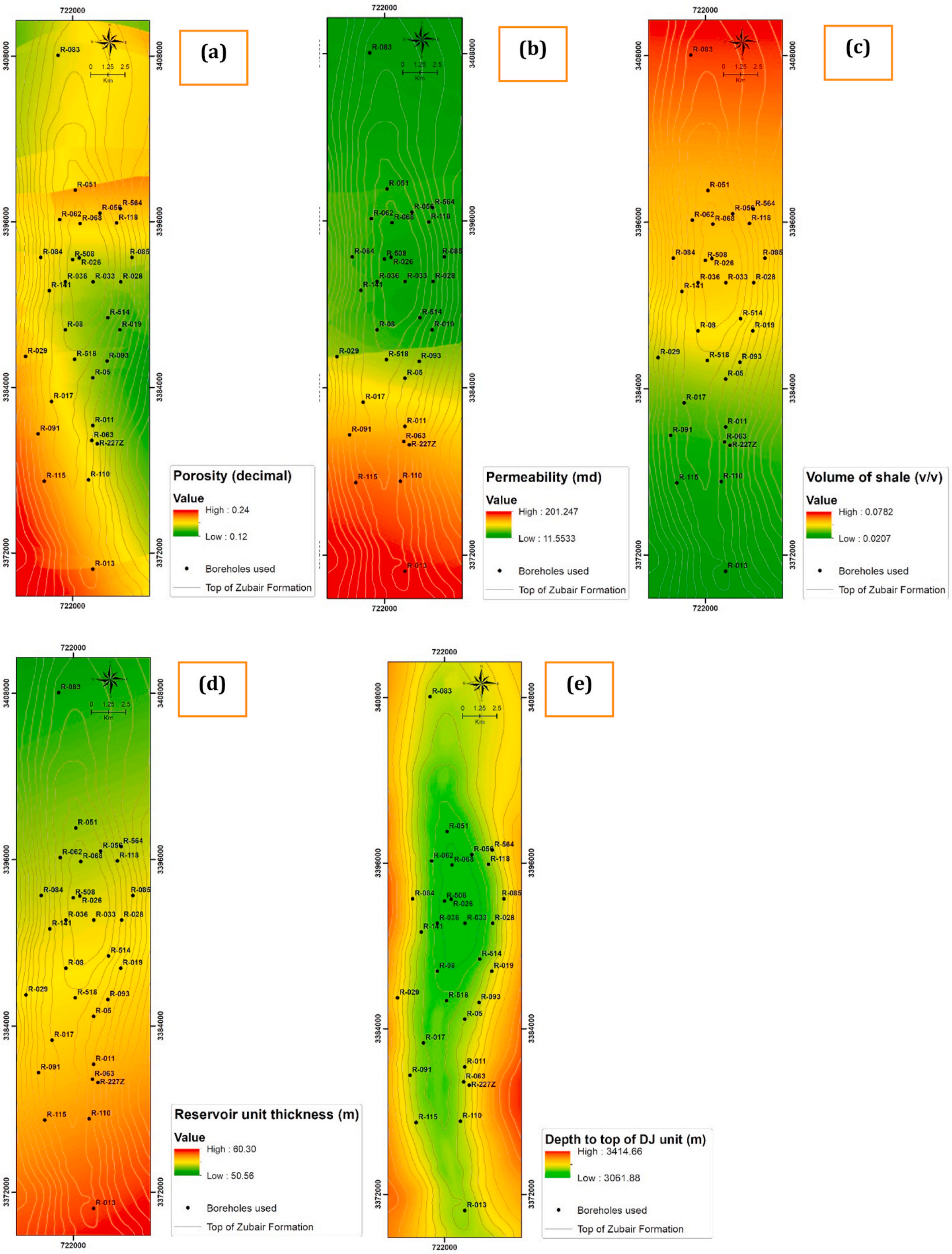


Fig. 7. Interpolated surfaces of factors affecting heterogeneity: (a) porosity (b) permeability (c) volume of shale (d) DJ unit thickness, and (e) depth to the top of DJ unit.

the root through the cost-complexity pruning process. The next split to be pruned is the one that least contributes to the overall output of the tree on training data. The mechanism of CART that produces an accurate predictive model depends on building a sequence of complex and overlapping trees rather than producing a single tree. By assessing the predictive performance of each tree in the pruning sequence based on independent test data, the preferred “right-sized” tree is determined. The CART is attractive because it models data in a way that is easy to visualize and interpret; the data used in the CART model can be of any type (numeric, binary, categorical, etc.); resistance to outliers is inherent; the handling of missing data is automatic; class balancing is automatic; and cost-sensitive, learning, dynamic-feature construction and probability-tree estimation are allowed (Aertsen et al., 2010; Breiman et al., 1984; Wu and Kumar, 2009).

4.4.2. BRT

BRT is an ensemble of powerful learning strategies aimed at enhancing a single model’s efficiency by fitting and integrating multiple models for prediction (Elith et al., 2008). BRT incorporates the strength of two algorithms, regression trees and boosting, to build a predictive model. Regression trees are models that relate a response to their predictors by recursive binary splits, whereas boosting is a strategy to combine the output of many “weak learners” into a powerful “committee” (Hastie et al., 2009). The outcome is an additive regression

Table 4
Variogram analysis of the EBK for the considered factors (bold refers to the best model).

Factor average	Semivariogram					
	without transformation		With transformation			
	Model	RMSE	Empirical	RMSE	Log Empirical	RMSE
φ (Arithmetic)	Linear	0.02261	Exponential	0.02222	Exponential	0.02223
	Power	0.02283	Exponential detrend	0.02147	Exponential detrend	0.02149
	Thin plate spline	**0.02092**	Whitte	0.02210	Whitte	0.02215
			Whitte detrend	0.02220	Whitte detrend	0.02214
			K-Bessel	0.02199	K-Bessel	0.02197
			K-Bessel detrend	0.02124	K-Bessel detrend	0.02121
φ (Geometric)	Linear	0.02155	Exponential	0.02178	Exponential	0.02184
	Power	0.02174	Exponential detrend	0.02113	Exponential detrend	0.02104
	Thin plate spline	0.02092	Whitte	0.02153	Whitte	0.02158
			Whitte detrend	0.02091	Whitte detrend	0.20957
			K-Bessel	0.02126	K-Bessel	0.02129
			K-Bessel detrend	0.02075	K-Bessel detrend	0.02077
K (Arithmetic)	Linear	260.91	Exponential	258.63	Exponential	259.34
	Power	260.37	Exponential detrend	242.01	Exponential detrend	248.60
	Thin plate spline	270.96	Whitte	257.36	Whitte	257.44
			Whitte detrend	242.15	Whitte detrend	248.72
			K-Bessel	255.18	K-Bessel	255.67
			K-Bessel detrend	242.53	K-Bessel detrend	249.37
K (Geometric)	Linear	79.93	Exponential	78.78	Exponential	78.88
	Power	78.55	Exponential detrend	73.15	Exponential detrend	75.09
	Thin plate spline	83.00	Whitte	77.37	Whitte	76.92
			Whitte detrend	74.03	Whitte detrend	74.95
			K-Bessel	76.01	K-Bessel	77.07
			K-Bessel detrend	**73.14**	K-Bessel detrend	74.44
V_{sh} (Arithmetic)	Linear	0.0021	Exponential	0.0025	Exponential	0.0025
	Power	0.0018	Exponential detrend	0.0019	Exponential detrend	0.0018
	Thin plate spline	0.0020	Whitte	0.0021	Whitte	0.0021
			Whitte detrend	0.0019	Whitte detrend	0.0018
			K-Bessel	0.0019	K-Bessel	0.0018
			K-Bessel detrend	0.0019	K-Bessel detrend	**0.0017**
Unit thickness (Arithmetic)	Linear	4.213	Exponential	4.137	Exponential	4.129
	Power	4.221	Exponential detrend	3.795	Exponential detrend	**3.762**
	Thin plate spline	4.787	Whitte	4.145	Whitte	4.148
			Whitte detrend	3.794	Whitte detrend	**3.762**
			K-Bessel	4.161	K-Bessel	4.137
			K-Bessel detrend	3.806	K-Bessel detrend	3.771

model trained in a forward, stage-wise manner. The major advantage of BRT compared to other decision tree models is its ability to deal with different types of data (numeric, binary, categorical, etc.) and not be affected by the missing data. Also, the data used as predictors to build a BRT model do not need to transform, standardize, or deleting the outliers (Al-Abadi and Al-Najar, 2020). BRT is also distinguished by its ability to easily fit non-linear models in addition to accounting for the effects of interactions among the predictors (Elith et al., 2008).

4.4.3. RF

RF is an ensemble, supervised, learning algorithm that combines the concept of DT and bagging to solve both regression and classification problems (Breiman, 2001). Bagging is a technique for producing multiple training data by resampling with the replacement of the original training set (Carranza et al., 2020). Bagging generates several DTs from resampled data and combines the predicted values through averaging and voting. During bagging, 1/3 of the data that are not utilized during tree construction, referred to as out-of-bag (OOB) observations, are used as test data to evaluate the misclassification-error rate and estimate predictive accuracy. The RF algorithm has the intrinsic potential to measure predictor importance by determining how far the prediction error increases when OOB observations are permuted for the predictor while other predictors remain unchanged. The RF is also capable of handling missing values, resistance to overfitting, and large datasets with higher dimensionality (Handhal et al., 2020a).

4.4.4. NB

The NB is a family of simple probabilistic algorithms based on Baye’s theorem with strong independence assumptions between the features (Shmueli et al., 2017). NB has numerous advantages, including ease of design and construction, requires no sophisticated estimate of iterative parameters, and robustness to irrelevant feature and noise (Soria et al., 2011).. The main weakness of NB lies in the assumption that the factors involved in model construction must be independent of each other; however, if there are factors dependent on each other, a large number of incorrect classifications may occur (Pham et al., 2017).

4.4.5. SVM

SVM is a supervised machine learning technique designed to solve both classification and regression problems. It seeks to find a hyperplane that best divides a dataset into distinguished classes. A hyperplane is a line that linearly separates a dataset (Pal and Mather, 2005). Support vectors, on the other hand, are the data points located close to the hyperplane. They are critical elements of the SVM as removing these points dramatically changing the position of the dividing hyperplane. The margin is the distance between the hyperplane and the support vectors. The SVM algorithm chooses a hyperplane with the greatest possible margin between the hyperplane and any point within the training dataset, and thus the new data could be classified correctly (Vapnik and Chervonenkis, 1974). In case of difficulty to define a clear hyperplane, the two-dimensional data is converted to a three-dimension through kernelling concept and the hyperplane turns into a plane. The dataset is continually mapped into higher and higher dimensions until a hyperplane can be formed for optimal segregate. The selection of kernel function and its parameters is crucial for successful application of SVM (Al-Mayahi et al., 2021). There are different types of kernel functions and their choice depends mostly on the nature and features of the examined phenomena (Arabgol et al., 2016). In this study, three kernel functions were selected: linear, radial, and polynomial. The math behind SVM can be elsewhere (Vapnik, 2013).

4.4.6. KNN

KNN is a non-parametrically supervised algorithm that can be used for both classification and regression problems. The KNN uses "similar" example in the training data to predict a new case. These "neighbors" are then used to predict the new instance by voting (for classification problem) or averaging (for regression) (Handhal et al., 2020a). Its simplicity, lake of parametric assumption, and robustness to noisy training data are the main advantage of this technique.

4.4.7. LR

LR is a classification algorithm that is used where the response variable has two only two possible outcomes (dichotomous variable) (Al-Abadi and Al-Najar, 2020). The idea behind this technique is to find a relationship between features and probability of particular outcome. The data used to build LR model do not need to be normally distributed and can be continuous, categorial, or both (Lee and Sambath, 2006). The algorithm of LR applies maximum likelihood estimation after transforming the dependent variable into a logit variable (Bai et al., 2010). The LR generates the coefficient of a formula to predict a logit transformation of the probability of the presence of the characteristic of interest.

4.5. Model performance evaluation

To evaluate the prediction accuracy of the classifiers used in this study, two statistical measures were used: accuracy and Cohen's kappa. Accuracy is the proportion of observations that are correctly classified. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FB + FN} \tag{4}$$

where TP is the number of boreholes predicted as positive (heterogeneous) that turn out to be positive, TN is the number of boreholes predicted as negative (homogeneous) that turn out to be negative, FP is the number of boreholes predicted as positive that turn out to be negative, and FN is the number of boreholes predicted as negative that turn out to be positive.

Cohen's kappa is a measure used to investigate the inter-rater reliability or agreement between two raters (Al-Abadi and Alsamaani, 2020). It can be calculated as:

$$k = \frac{\pi_0 - \pi_e}{1 - \pi_0} \tag{5}$$

where π_0 is an observational probability of agreement, and π_e is a hypothetical expected probability of agreement under an appropriate set of baseline constraints, such as total independence of observer classifications (Landis and Koch, 1977). The predictive model is said to be slight if k is between 0.01 and 0.20, fair if k is between 0.21 and 0.40, moderate if k is between 0.41 and 0.60, substantial if k is between 0.61 and 0.80, and an almost perfect if k is between 0.80 and 1.00.

4.6. Software used for implementing classifiers

The three machine-learning models used in this study were implemented using the Caret package (Kuhn, 2008) in R software. Caret is a group of functions that aim to streamline the predictive model-development process. Caret offers many tools for data splitting, pre-processing, feature selection, model tuning, variable importance evaluation, and many others.

5. Results and discussion

5.1. Data used

The values of core φ data range from 0.001 to 0.283 with an average of 0.174, 0.056 standard deviation, and 32.33% coefficient of variation. In contrast, the core k values vary from 0 to 6405 md with an average of 384.4, 644.4 standard deviation, and 167.6% coefficient of variation. Overall, both parameters are skewed, but φ exhibits less variation than k. The probability distributions of φ and log(k) are approximately normal (Fig. 5). The estimated φ values from the NMR log data are between 0.008 and 0.298 with an average of 0.298 and a standard deviation equal to 0.017. The coefficient of variation is 29.53%. Concerning k values derived from the SDR model, the range of these values is from 0 to 5568 with an average of 280.3 m. The standard deviation and coefficient of variation are 731.6 and 261.05%, respectively. The probability distributions of φ and log(k) are also approximately normal (Fig. 5). The estimated k values from the SDR model indicate an acceptable fit with the k data obtained from the core laboratory (Table 1).

Results of regression and correlation analyses (Table 2) proved that there is a strong positive relationship between log(k) and φ (correlation of determination equal to 0.68) that improved by the addition of V_{sh} as a factor influencing the k distribution ($R^2 = 0.74$) (multiple-regression model). This hold true for the reservoir unit's homogeneous interval in the southern portion of the oil field, where the R-013 well is located. The correlation between log(k) and V_{sh} is negative, implying that as V_{sh} increases, k decreases. The connection between log(k) and φ and V_{sh} is weak to moderate in the central and northern portions of the oil field, where the R-508 and R-083 wells are located. In general, the relationship between log(k) and φ is positive, whereas the relationship between log(k) and Vsh is negative, indicating that the spatial k and φ distribution of a sedimentary unit can be used to spatially predict vertical heterogeneity of a reservoir unit in the study area.

5.2. Select the best variogram

From Table 4, it is obvious that *thin plate spline* semivariogram model for the arithmetic φ give the best performance model with lowest RMSE (equal to 0.02092) compared with other semivariograms (with and without transformation). On the other hand, the K-Bessel detrend (empirical transformed semivariogram) is the best one to simulate the k geometric mean across the reservoir unit with RMSE equal to 73.14. In the case of V_{sh}, the best model was log Empirical k-Bessel detrend with RMSE equal to 0.0017. The other models were good too either without and with the transformation process. Finally, both log Empirical exponential detrend and Whitte detrend are suitable to represent the semivariogram of the DJ unit reservoir thickness with RMSE equal to 3.762 for both models.

5.3. Spatial distribution of factors affecting heterogeneity

Through the spatial distribution of the average φ values in the field (Fig. 7a), it is clear that the φ in southern parts of the study area is generally greater than the φ values in the rest of the field. Smaller values of this petrophysical parameter generally appear in the middle, and moderate values generally occur in the northern parts of the field. In contrast, high k values (Fig. 7b) were recorded in the southern parts of the reservoir unit, and low values appear to the north. Moderate k values generally occupy the middle parts of the field. For the V_{sh} values, they show a different spatial distribution to that of φ and k. The V_{sh} distribution in the study area (Fig. 7c) shows a different pattern for the distribution of φ and k, as the volume of shale increases from south to north. This distribution of V_{sh} reflects the effect of this factor on both φ and k, which in turn reduces the values of these two petrophysical parameters. The thickness of the reservoir unit (Fig. 7d) generally increases from north to south with moderate thickness values occurring in the middle of the field. Finally, as the Rumaila oil field is an anticline, the high depths to the reservoir top appear on the flanks of the fold, and low depths to the top of the reservoir appear along the fold axis (Fig. 7e).

5.4. Training and validating classifiers

The training results of the seven machine classifiers using in this study, in terms of accuracy and Cohen's kappa, were presented in Table 5. The random-search technique was used to optimize the hyperparameters of each classifier. From Table 5, the RF classifier showed the highest accuracy (0.991), followed by SVM-radial (0.983), BRT (0.960), LR (0.955), and CART (0.944). The other classifiers are performed well too in the training stage (accuracy >0.80). In terms of Cohen's kappa, all classifiers had almost perfect performance (>0.8), except KNN, Naïve Bayes, and SVM-polynomial which had substantial performance (between 0.61 and 0.8). The best performance of the models in terms of Cohen's kappa is RF (0.970) followed by SVM-radial (0.966), SVM-Linear (0.915), BRT (0.912), LR (0.911), and CART (0.880). Investigating the importance of variables for the tree-based classifiers (CART, BRT, and RF) proved that the most important factors in building the classifier models were Vsh, k, and φ (Fig. 8). The less important factors were unit thickness and depth to the top of the reservoir unit. Overall, all factors play a role in controlling the distribution of heterogeneity across the oil field. After training the classifiers was successful, the test dataset was passed to each classifier, and the results were compared (Table 5). The highest classification accuracy also belonged to RF (0.982), followed by SVM-radial (0.966), LR (0.944), BRT (0.865), CART (0.825), and Naïve Bayes (0.820). The lowest classification accuracy belonged to SVM-polynomial (0.730) and KNN (0.761). In terms of Cohen's kappa, the almost perfect models were RF (0.962), followed by SVM-radial (0.960), LR (0.881), and SVM-linear (0.880). The CART, BRT, and Naïve Bayes were substantial significance and the KNN and Naïve Bayes were moderate performance models.

Examine the results (Table 5) proved that RF, SVM-radial, and LR performed much better than others in both training and testing stages phases. Therefore, the findings of these models were chosen to show the spatial heterogeneity of the DJ unit in the study area.

5.5. Mapping the probability of spatial heterogeneity

The probability values of the three classifiers (RF, SVM-radial, and LR) for training and testing phases were exported to ArcGIS 10.7 software, interpolated using EBK, and then visualized using five classes: very low, low, moderate, high, and very high (Fig. 9a–c). The natural-break classification scheme (Jenks, 1967) was used to classify probability values into different categories. The natural break is an optimal classification method that seeks out class breaks that minimize within-class variance and maximize between-class differences. The areas occupied by the five classes were presented in Table 6. For the RF and SVM-radial models, the high-very high classes occupy ~60% (274 km^2) of the study area and mainly distribute over middle and northern parts of the field. The very low-to-low zones encompass ~30% (142 km^2) and occur in southern parts of the field. A zone of moderate heterogeneity is only present in a small strip between the low and high zones and occupies only 10% (53 km^2) of the field area (Fig. 9). For the LR model, there is a little difference between the areas occupied by the heterogeneity zones and that encompass by the RF and SVM-radial, the very low-low, moderate, and high-very high occupy 29% (134 km^2), 7% (35 km^2), and 64% (300 km^2), respectively. The spatial distribution of heterogeneity for this model is similar to the other two models; high values in northern parts and low values in the southern portions.

Table 5
Evaluation of the classifiers' performances. Bold refers to the best models.

Model	Training		Testing	
	Accuracy	Cohen's kappa	Accuracy	Cohen's kappa
CART	0.944	0.880	0.825	0.650
BRT	0.960	0.912	0.865	0.711
RF	**0.991**	**0.970**	**0.982**	**0.962**
SVM-Linear	0.955	0.915	0.944	0.880
SVM-Radial	**0.983**	**0.966**	**0.980**	**0.960**
SVM-Polynomial	0.815	0.628	0.730	0.500
Naïve Bayes	0.865	0.734	0.820	0.690
KNN	0.880	0.761	0.790	0.584
LR	**0.955**	**0.911**	**0.944**	**0.881**

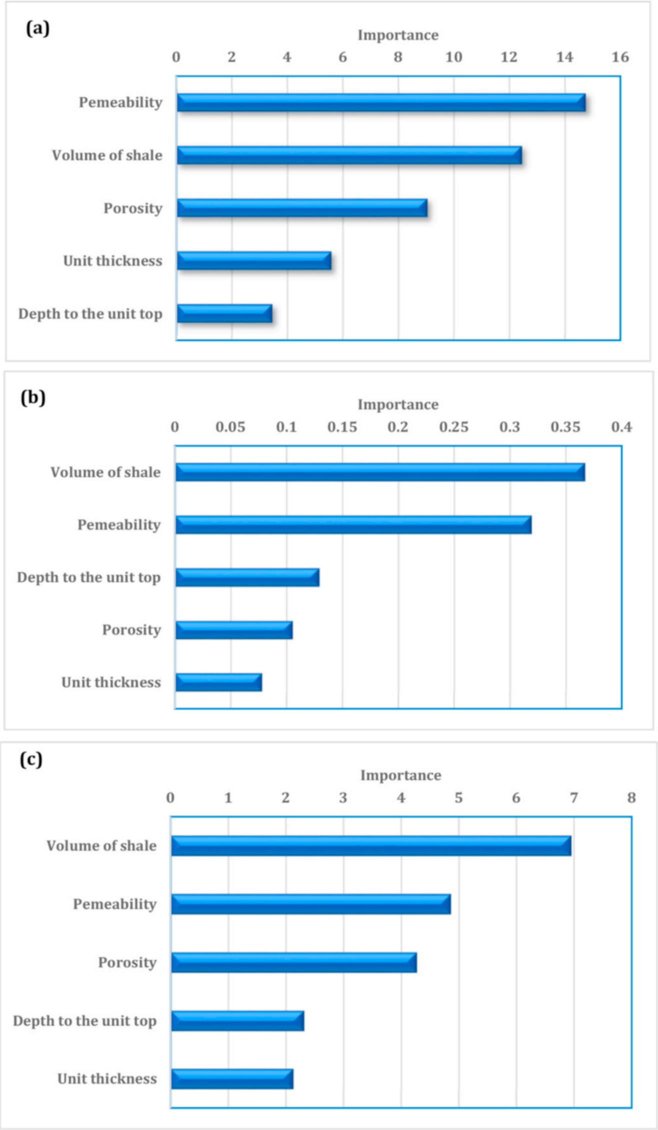


Fig. 8. Variable importance based on: (a) CART (b) BRT (c) RF.

The mapping of heterogeneity (Fig. 9) and those factors influencing heterogeneity (Fig. 7) show that the mapped zones bear a clear relationship to depositional environments. The Zubair Formation has been interpreted to represent the eastward and northeastward progradation of a delta system into a deeper-water, tide-influenced estuary to the east and northeast (Jassim and Goff, 2006). With source areas to the west and south on the Arabian Shield, sands were apparently transported to the estuary via a river system controlled by the Al-Batin transverse fault system (Aqrawi et al., 2010). Although some of the sands were deposited in very nearshore lagoonal to marsh areas, most of the sand was transported farther seaward into delta-front, distributary-channel environments, where the sands were reworked into well-sorted, cross-bedded,

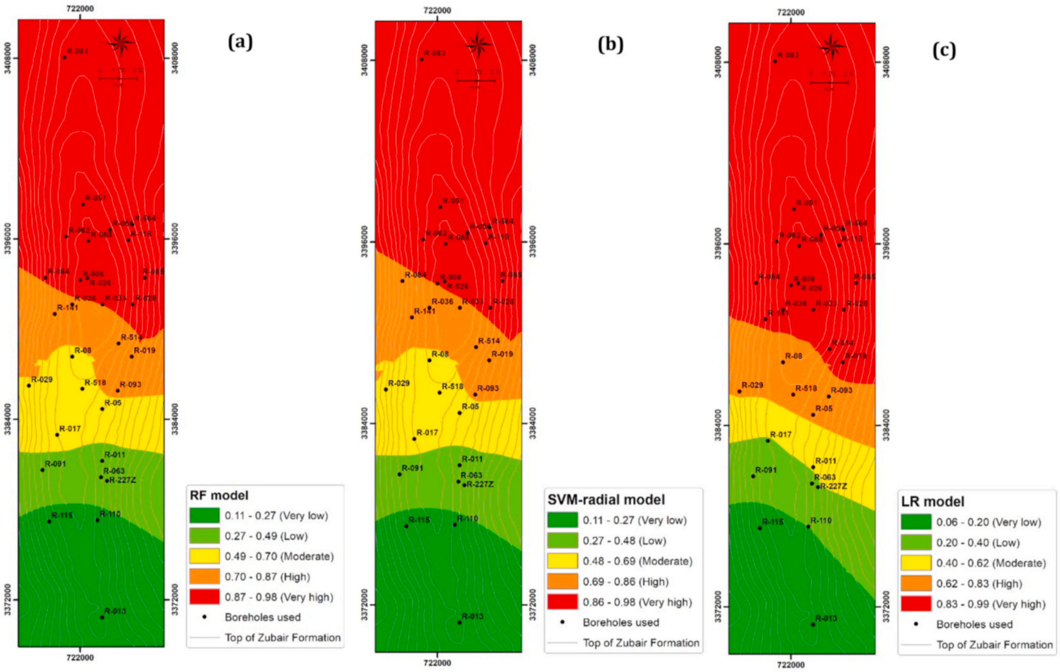


Fig. 9. Spatial distribution of the probability of heterogeneity in the DJ unit (a) RF (b) SVM-radial, (c) LR.

Table 6
Areas occupy by heterogeneity classes for the best three classifiers.

GP zone	RF		SVM-radial		LR	
	Area (%)	Area (km^2)	Area (%)	Area (km^2)	Area (%)	Area (km^2)
Very low	0.203	95	0.202	95	0.186	87
Low	0.098	46	0.101	47	0.099	47
Moderate	0.109	51	0.116	55	0.074	35
High	0.112	52	0.125	58	0.104	49
Very high	0.478	224	0.456	214	0.537	252

fine-to medium-grained arenites by waves and tidal currents (e.g., Al-Zaidy (2020)). It is these sand bodies, like the DJ unit of the upper sandstone member, which form the well-sorted, pure quartz arenites of the main pay zone. They are concentrated in the southern end of the Rumaila field because that is where the rivers initially debouched their loads into the estuary to be reworked. Finer-grained sands, silts and muds were transported farther seaward to the north and east into deeper-water, prodeltaic parts of the estuary, where they became interbedded sandstones and mudstones, siltstones, and shales, explaining the distribution of heterogeneity and influencing factors shown in Figs. 9 and 7. For example, the unit thickness is greater in the south of the field than in the northern parts (Fig. 7d), reflecting the fact that southern parts of the field are closer to proximal source areas. Similarly, the porosity and permeability are also greater in southern parts of the field (Fig. 7a and b), because this area represents more proximal, higher-energy parts of the former delta where winnowing of the fines left behind pure quartz arenites with high porosities and permeabilities. Moreover, irregular sub-horizontal to sub-vertical bands of increased thickness and porosity to the north (Fig. 7a and d) probably represent thicker, purer sands in highly reworked offshore bars and distributary channels. Overall, however, the volume of shales in the DJ unit increases to the north (Fig. 7c) in more distal, deeper-water portions of the delta, and because of the increase in the volume of finer-grained rocks, overall heterogeneity increases to the north as well (Fig. 9).

6. Conclusions

The main conclusions of this study are: (1) the integration of the Lorenz heterogeneity index, geostatistics, and machine-learning classifiers can provide a simple and easy way to study and predict the spatial distribution of vertical heterogeneity of a reservoir unit. (2) The RF, SVM-radial, and LR classifiers were more powerful than other algorithms used in this study in modeling the spatial heterogeneity of the reservoir unit being studied. (3) The spatial volume of shale distribution can be correlated to the vertical heterogeneity of the DJ unit, and (4) The DJ reservoir unit is more heterogeneous in the middle and northern parts of the field than in southern parts. The fact that gross formation heterogeneity and influencing factors in the DJ unit mirror closely former depositional environments across the Rumaila field, suggests the validity of the methods in that they reflect actual depositional processes. Clearly, the use of geostatistics and GIS-based machine-learning classifiers show how effective the methods described herein can be the in the analysis of limited field data.

Credit author statement

Amna M. Handhal: Conceptualization, Methodology, Data curation. Frank R. Ettensohn: Supervision, Writing-Reviewing and Editing. Alaa M. Al-Abadi: Supervision, Visualization, Software, Writing-Reviewing and Editing. Maher J. Ismail: Methodology, Visualization, Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to express our heartfelt gratitude to the editorial board and the respected reviewers for their tireless efforts in revising the manuscript, particularly reviewer #2, without tireless efforts in revising the article draft, this study would not exist in its current form.

References

Aertsen, W., Kint, V., Van Orshoven, J., Özkan, K., Muys, B., 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. Ecol. Model. 221, 1119–1130.

Ahmed, T., 2018. Reservoir Engineering Handbook. Gulf professional publishing.

Al-Abadi, A.M., Al-Najar, N.A., 2020. Comparative assessment of bivariate, multivariate and machine learning models for mapping flood proneness. Nat. Hazards 100, 461–491.

Al-Abadi, A.M., Alsamaani, J.J., 2020. Spatial analysis of groundwater flowing artesian condition using machine learning techniques. Groundw. Sustain. Dev. 100418.

Al-Abadi, A.M., Handhal, A.M., Al-Ginamy, M.A., 2019. Evaluating the dibdibba aquifer productivity at the karbala–najaf plateau (Central Iraq) using GIS-based tree machine learning algorithms. Nat. Resour. Res. 1–21.

Al-Ansari, R., 1993. The Petroleum Geology of the Upper Sandstone Member of the Zubair Formation in the Rumaila South. Geol. Study. Minist. Oil, Baghdad, Iraq.

Al-Mayahi, H.M., Al-Abadi, A.M., Fryar, A.E., 2021. Probability mapping of groundwater contamination by hydrocarbon from the deep oil reservoirs using GIS-based machine-learning algorithms: a case study of the Dammam aquifer (middle of Iraq). Environ. Sci. Pollut. Res. 28, 13736–13751.

Al-Mudhafar, W.J., 2019. Bayesian kriging for reproducing reservoir heterogeneity in a tidal depositional environment of a sandstone formation. J. Appl. Geophys. 160, 84–102.

Al-Mudhafar, W.J., Hakim, S., 2015. Parallel programming of model-based geostatistics for improved reservoir characterization. Int. Assoc. Math. Geosci. Freiberg, Ger.

Al-Zaidy, A.A.H., 2020. Facies architecture and stratigraphic sequence of Zubair Formation in majnoon and suba oil fields, southern Iraq. Model. Earth Syst. Environ. 6, 779–792.

Almalikee, H.S., Al-Najm, F.M., 2019. Wellbore stability analysis and application to optimize high-angle wells design in Rumaila oil field, Iraq. Model. Earth Syst. Environ. 5, 1059–1069.

Alshayef, M.S., Javed, A., Mohammed, A.M. Bin, 2019. Delineation of hydrocarbon potential zones in Masila oil field. Yemen. Spat. Inf. Res. 27, 121–135. https://doi.org/10.1007/s41324-018-0220-0.

Amiri, M.A., Karimi, M., Sarab, A.A., 2015. Hydrocarbon resources potential mapping using evidential belief functions and frequency ratio approaches, southeastern Saskatchewan, Canada. Can. J. Earth Sci. 52, 182–195. https://doi.org/10.1139/cjes-2013-0193.

Aqrawi, A.A.M., Goff, J.C., Horbury, A.D., Sadooni, F.N., 2010. The Petroleum Geology of Iraq. Scientific Press.

Arabgol, R., Sartaj, M., Asghari, K., 2016. Predicting nitrate concentration and its spatial distribution in groundwater resources using support vector machines (SVMs) model. Environ. Model. Assess. 21, 71–82.

Bai, S.-B., Wang, J., Lü, G.-N., Zhou, P.-G., Hou, S.-S., Xu, S.-N., 2010. GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. Geomorphology 115, 23–31.

Bellen, R.C. Van, Dunnington, H.V., Wetzel, R., Morton, D., 1959. Lexique Stratigraphique International, vol. 3. Asie, Iraq.

Bingham, L., Zurita-Milla, R., Escalona, A., 2012. Geographic information system-based fuzzy-logic analysis for petroleum exploration with a case study of northern South America. Am. Assoc. Petrol. Geol. Bull. 96, 2121–2142. https://doi.org/10.1306/04251212009.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933 404324.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC press.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Network. 106, 249–259.

Carranza, C., Nolet, C., Pezij, M., Van Der Ploeg, M., 2020. Root zone soil moisture estimation with Random Forest. J. Hydrol 593, 125840.

Chawla, N.V., 2009. Data mining for imbalanced datasets: an overview. Data Mining and Knowledge Discovery Handbook. Springer, pp. 875–886.

De Ros, L.F., 1998. Heterogeneous generation and evolution of diagenetic quartzarenites in the silurian-devonian furnas formation of the paraná basin, southern Brazil. Sediment. Geol. 116, 99–128.

Diggle, P., Lophaven, S., 2006. Bayesian geostatistical design. Scand. J. Stat. 33, 53–64.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.

Fitch, P.J.R., Lovell, M.A., Davies, S.J., Pritchard, T., Harvey, P.K., 2015. An integrated and quantitative approach to petrophysical heterogeneity. Mar. Petrol. Geol. 63, 82–96.

Fouad, S.F.A., 2010. Tectonic and structural evolution of the mesopotamia foredeep, Iraq. Iraqi Bull. Geol. Min. 6, 41–53.

Fouad, S.F.A., Sissakian, V.K., 2011. Tectonic and structural evolution of the mesopotamia plain. Iraqi Bull. Geol. Min. 33–46.

Frazer, G.W., Wulder, M.A., Niemann, K.O., 2005. Simulation and quantification of the fine-scale spatial pattern and heterogeneity of forest canopy structure: a lacunarity-based method designed for analysis of continuous canopy heights. For. Ecol. Manage. 214, 65–90.

Handhal, A.M., Al-Abadi, A.M., Chafeet, H.E., Ismail, M.J., 2020a. Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. Mar. Petrol. Geol. 104347.

Handhal, A.M., Hussein, A.A., Al-Abadi, A.M., Ettensohn, F.R., 2020b. Spatial modeling of hydrocarbon productivity in the Nahr Umr Formation at the Luhais oil field, southern Iraq. Nat. Resour. Res. 1–23.

Handhal, A.M., Jawad, S.M., Al-Abadi, A.M., 2019. GIS-based machine learning models for mapping tar mat zones in upper part (DJ unit) of Zubair Formation in North Rumaila supergiant oil field, southern Iraq. J. Petrol. Sci. Eng. 178 https://doi.org/10.1016/j.petrol.2019.03.071.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Jaffar, H.M., 2018. Structural Geology of Rumaila Oilfield in Southern Iraq from Well Logs and Seismic Data. Unpub. M. Sc. theses. Univ. Basrah.

Jaffar, H.M., Abdulnaby, W., 2018. Stress regime of rumania oilfield in southern Iraq from borehole breakouts. IOSR J. Appl. Geol. Geophys. 6, 25–35.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intell. Data Anal. 6, 429–449.

Jassim, S.Z., Goff, J.C., 2006. Geology of Iraq. DOLIN, Sro, Distributed. Geological Society of London.

Jenks, G.F., 1967. The data model concept in statistical mapping. Int. Yearb. Cartogr. 7, 186–190.

Kelkar, M., Perez, G., Chopra, A., 2002. Applied Geostatistics for Reservoir Characterization. Society of Petroleum Engineers Richardson, TX.

Krivoruchko, K., 2011. Spatial Statistical Data Analysis for GIS Users. Esri Press Redlands.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Software 28, 1–26.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 159–174.

Lawrence, S., Burns, I., Back, A., Tsoi, A.C., Giles, C.L., 1998. Neural network classification and prior class probabilities. Neural Networks: Tricks of the Trade. Springer, pp. 299–313.

Lee, S., Sambath, T., 2006. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environ. Geol. 50, 847–855.

Liu, X., Zhong, G., Yin, J., He, Y., Li, X., 2008. GIS-based modeling of secondary hydrocarbon migration pathways and its application in the northern Songliao Basin, northeast China. Comput. Geosci. 34, 1115–1126. https://doi.org/10.1016/j.cageo.2007.08.005.

Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D., 2008. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Network. 21, 427–436.

Morad, S., Al-Ramadan, K., Ketzer, J.M., De Ros, L.F., 2010. The impact of diagenesis on the heterogeneity of sandstone reservoirs: a review of the role of depositional facies and sequence stratigraphy. Am. Assoc. Petrol. Geol. Bull. 94, 1267–1309.

Nowak, W., De Barros, F.P.J., Rubin, Y., 2010. Bayesian geostatistical design: task-driven optimal site investigation when the geostatistical model is uncertain. Water Resour. Res. 46.

Owen, R.M.S., Nasr, S.N., 1958. Stratigraphy of the Kuwait-Basra Area: Middle East.

Pal, M., Mather, P.M., 2005. Support vector machines for classification in remote sensing. Int. J. Rem. Sens. 26, 1007–1011.

Pham, B.T., Bui, D.T., Dholakia, M.B., Prakash, I., Pham, H.V., Mehmood, K., Le, H.Q., 2017. A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS. Geomatics, Nat. Hazards Risk 8, 649–671.

Ren, H.-J., Wang, X.-C., Guo, Q.-L., Guo, X.-X., Zhang, R., 2020. Spatial prediction of oil and gas distribution using Tree Augmented Bayesian network. Comput. Geosci. 104518.

Rivoirard, J., 2005. Concepts and methods of geostatistics. Space, Structure and Randomness. Springer, pp. 17–37.

Rudini, Matori, A.N., Ab Talib, J., Balogun, A.L., 2018. Application of geographic information system (GIS) to model the hydrocarbon migration: case study from north-east Malay basin, Malaysia. E3S Web Conf 34. https://doi.org/10.1051/e3sconf/20183402027.

Shmueli, G., Bruce, P.C., Yahav, I., Patel, N.R., Lichtendahl Jr., K.C., 2017. Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. John Wiley & Sons.

Soria, D., Garibaldi, J.M., Ambrogi, F., Biganzoli, E.M., Ellis, I.O., 2011. A 'non-parametric'version of the naive Bayes classifier. Knowl. Base Syst. 24, 775–784.

Tiab, D., Donaldson, E.C., 2015. Petrophysics: Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties. Gulf professional publishing.

Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer science & business media.

Vapnik, V., Chervonenkis, A., 1974. Theory of Pattern Recognition.

Wu, X., Kumar, V., 2009. The Top Ten Algorithms in Data Mining. CRC press.

Yarmohammadi, S., Kadkhodaie, A., Hosseinzadeh, S., 2020. An integrated approach for heterogeneity analysis of carbonate reservoirs by using image log based porosity distributions, NMR T2 curves, velocity deviation log and petrographic studies: a case study from the South Pars gas field, Persian Gulf Basin. J. Petrol. Sci. Eng. 107283.

Zhou, Z.-H., Liu, X.-Y., 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. 18, 63–77.