energies

MDPI

Article

Improving Subsurface Characterisation with ‘Big Data’ Mining and Machine Learning

Rachel E. Brackenridge [1,*], Vasily Demyanov [2], Oleg Vashutin [3] and Ruslan Nigmatullin [3]

[1] School of Geosciences, University of Aberdeen, Aberdeen AB24 3FX, UK
[2] School of Energy, Geoscience, Infrastructure and Society, Heriot Watt University, Riccarton, Edinburgh EH14 4AS, UK; v.demyanov@hw.ac.uk
[3] Wood Mackenzie, 125009 Moscow, Russia; Oleg.Vashutin@woodmac.com (O.V.); Ruslan.Nigmatullin@woodmac.com (R.N.)
* Correspondence: Rachel.Brackenridge@abdn.ac.uk

Abstract: Large databases of legacy hydrocarbon reservoir and well data provide an opportunity to use modern data mining techniques to improve our understanding of the subsurface in the presence of uncertainty and improve predictability of reservoir properties. A data mining approach provides a way to screen dependencies in reservoir and fluid data and enable subsurface specialists to estimate absent properties in partial or incomplete datasets. This allows for uncertainty to be managed and reduced. An improvement in reservoir characterisation using machine learning results from the capacity of machine learning methods to detect and model hidden dependencies in large multivariate datasets with noisy and missing data. This study presents a workflow applied to a large basin-scale reservoir characterization database. The study aims to understand the dependencies between reservoir attributes in order to allow for predictions to be made to improve the data coverage. The machine learning workflow comprises the following steps: (i) exploratory data analysis; (ii) detection of outliers and data partitioning into groups showing similar trends using clustering; (iii) identification of dependencies within reservoir data in multivariate feature space with self-organising maps; and (iv) feature selection using supervised learning to identify relevant properties to use for predictions where data are absent. This workflow provides an opportunity to reduce the cost and increase accuracy of hydrocarbon exploration and production in mature basins.

Keywords: big data; unsupervised learning; supervised learning; multivariant analysis; machine learning; hydrocarbon exploration; reservoir; subsurface characterisation

check for updates

Citation: Brackenridge, R.E.; Demyanov, V.; Vashutin, O.; Nigmatullin, R. Improving Subsurface Characterisation with ‘Big Data’ Mining and Machine Learning. *Energies* **2022**, *15*, 1070. https://doi.org/10.3390/en15031070

Academic Editors: Reza Rezaee and Galih Bangga

Received: 7 August 2021
Accepted: 18 January 2022
Published: 31 January 2022

1. Introduction

Since the pioneering oil wells of the mid to late 1800s [1], our knowledge and understanding of the subsurface has increased exponentially. Although geological and engineering knowledge has advanced with increasingly robust data acquisition and interpretation techniques, there remains an opportunity for data scientists to use modern data mining techniques on large databases of legacy hydrocarbon reservoir and well data to improve our understanding of the subsurface and increase exploration and production efficiency in hydrocarbon basins [2–4].

Data mining provides an objective mathematical way to identify hidden dependencies in ‘big’ datasets [5]. It has the capacity to process large multivariant databases to identify subtle dependencies that may otherwise be overlooked by manual inspection. Data mining unites a wide selection of data-driven algorithms capable of solving different types of problems: clustering, classification, regression, or probability estimation [6]. Data mining with analytical and machine learning methods can be used to address a wide range of practical problems that arise when dealing with hydrocarbon asset data to help inform management decisions.

There are many benefits to using data mining and machine learning techniques on legacy data, however, the hydrocarbon industry is yet to fully utilise 'big data' analytics [4]. This is despite the wide use of regression to predict reservoir and PVT (pressure-volume-temperature) values in the subsurface (e.g., [7–11]). Estimation of reservoir fluid properties benefit field development and production by informing: (i) production strategy to increase efficiency; (ii) enhanced oil recovery (EOR) planning; and (iii) production facilities design [9]. Formation Volume Factor (FVF) is particularly difficult and costly to measure or calculate. Multiple studies have proposed correlations between PVT data and FVF using gas/oil ratio (GOR), gas and oil gravity (API), and reservoir temperature and pressure (e.g., [8–11]). There is much uncertainty in FVF estimation from linear and non-linear regression or from graphical correlation methods [11]. In part, this uncertainty is driven by local variations in crude composition [7]. In addition, inputs to the correlation equation can be challenging to estimate. For example, gas gravity is rarely recorded [7], and GOR is commonly estimated from correlation equations where downhole or laboratory measurements have not been possible [12]. In recent years, a number of studies have tested the use of artificial neural networks (ANN) for PVT estimations [13–15]. These authors argue that ANN methods are superior to graphical or linear/non-linear multiple regression techniques by increasing estimation accuracy and reducing the limitations of correlations due to local crude composition differences. More recently, Oloso et al. [16] used neural networks to predict viscosity and GOR over a range of pressure values, and to account for differences in PVT data from basin to basin. Other studies have assessed the use of ANN and fuzzy logic [17] for permeability estimation [18–21]; reservoir properties and lithofacies prediction using ANN and Committee Machines [22]; and predicting Recovery Factor using Bagging and Random Forest methods [23]. ANN and fuzzy logic have also been used in datasets from other extractive industries such as mining [24]. It is clear that machine learning provides an excellent opportunity to further our understanding of the subsurface.

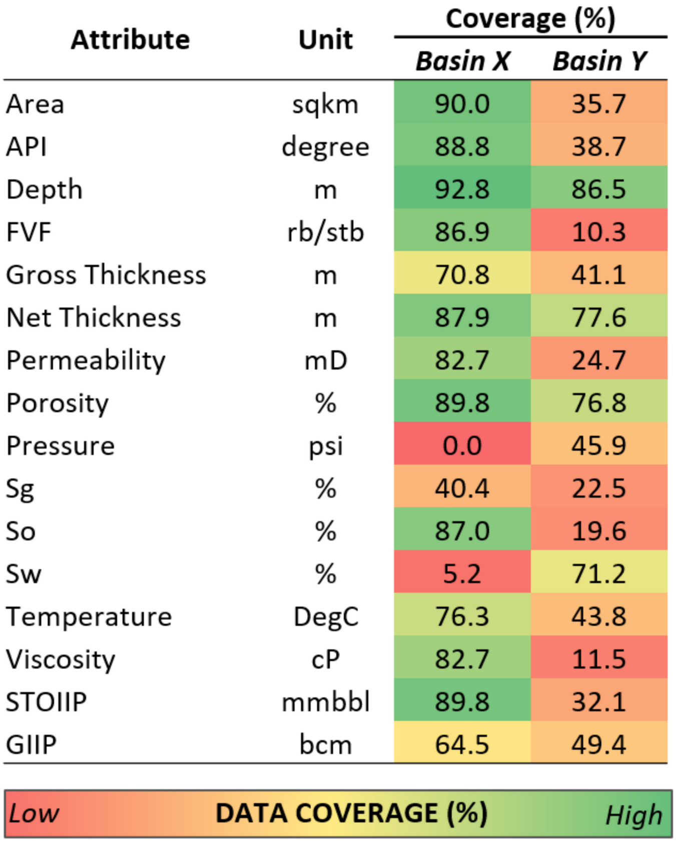
We propose that the use of data mining workflows on reservoir data can: (i) be used to rapidly quality-control and standardise large multivariant datasets. Following 50+ years of modern oil and gas exploration across multiple countries with different tools, measurement units and protocols, subsurface specialists require an efficient workflow to integrate datasets; (ii) reduce risk and uncertainty. Where information is lacking, a company may quantify the range of likely values (P10, P50, P90) in order to account for uncertainty. Alternatively, more data can be gathered, often at significant cost. Data mining and machine learning techniques can provide a cost-effective way to increase data coverage through predictions in legacy datasets where reservoir property data are partial or absent; and (iii) aid decision making. Strategic operational and economic decisions are routinely based on basic analysis and expert opinion of those datasets available to a company. Expert opinion is therefore readily biased by data selection and availability. Automated data mining and machine learning removes human bias and allows experts to focus on key decision drivers that are identified in the data.

The aim of this paper is to demonstrate a consistent data mining workflow to tackle the prediction of reservoir and fluid characteristics. This will be done using real-world datasets from two mature hydrocarbon basins. The workflow addresses the following aims: (i) assess the adequacy of the data by streamlining the process of outlier detection, error detection and identifying natural artifacts in the data; (ii) identify dependencies within the data; and (iii) improve data coverage and continuity by estimating missing data values. This study develops a workflow that could reduce risk and uncertainty, and aid decision making over the complete hydrocarbon value chain through exploration, appraisal, development and production. In addition, there is scope for legacy data and learnings from hydrocarbon basins to be re-analysed using machine learning to support the development of emerging energy technologies, such as subsurface storage, carbon storage and sequestration, and geothermal energy.

2. Data, Workflow and Methods

2.1. Data

Large legacy reservoir datasets from two mature hydrocarbon basins were analysed: *Basin X* and *Basin Y*. The data consists a variety of attributes collected from specific depths within reservoirs of various ages and depositional environments. Attributes document details of the reservoir properties and their hydrocarbon fill. In total, 14 attributes have been used for this study relating to the field (area), reservoir (depth; gross thickness; net thickness; permeability; porosity; temperature) and fluid fill (API; FVF; pore-fluid pressure; saturation of gas, oil and water (S_g, S_o, S_w); and viscosity) (Figure 1).



Attribute	Unit	Coverage (%)	
		Basin X	*Basin Y*
Area	sqkm	90.0	35.7
API	degree	88.8	38.7
Depth	m	92.8	86.5
FVF	rb/stb	86.9	10.3
Gross Thickness	m	70.8	41.1
Net Thickness	m	87.9	77.6
Permeability	mD	82.7	24.7
Porosity	%	89.8	76.8
Pressure	psi	0.0	45.9
Sg	%	40.4	22.5
So	%	87.0	19.6
Sw	%	5.2	71.2
Temperature	DegC	76.3	43.8
Viscosity	cP	82.7	11.5
STOIIP	mmbbl	89.8	32.1
GIIP	bcm	64.5	49.4

Low
DATA COVERAGE (%)
High

Figure 1. Attributes used in this study and their % data coverage across *Basin X* and *Basin Y*.

Basin X is a large (over 900 km^3) mature basin. The basin is predominantly oil-prone with a good mix of carbonate and clastic reservoirs. There are between 5 and 10 distinct plays in the basin that are largely controlled by structural domains: (a) deep Devo-Carboniferous rift plays; (b) Mid Carboniferous-Mid Triassic orogenic thrust belt and associated foreland basin plays; and (c) a longstanding carbonate-dominated platform domain. These structural trends and rift basin depocentres could exert some control on the distribution of source rock and reservoir properties across the region, although we note that a number of plays are spatially overlapping. *Basin X* consists of 10,557 unique point source datasets that have been gathered from the reservoir. Data coverage is good (generally over 80%) with notable exceptions for pressure and water saturation data where many values are absent (Figure 1). In total, 6,392 points have values for all the remaining attributes (area, API, depth, FVF, gross thickness, net thickness, permeability, porosity, oil saturation, temperature and viscosity) and have been used to test the workflow presented in this study.

Basin Y is also a well-explored and producing basin. Unlike *Basin X*, it contains both oil and gas fields. The basin is in excess of 130,000 km^3. This Permo-Carboniferous to Triassic basin consists of clastic reservoirs at various play levels. It is tectonically comparable to

Basin X, with Permo-Carboniferous rift and sag deposits controlling the distribution of petroleum system elements. The *Basin Y* dataset is smaller, consisting of 908 points. Data coverage is lower than that seen in *Basin X*: generally, attributes have below 50% of values available (Figure 1). However, some attributes show better coverage compared to the *Basin X* dataset (for example, pore-fluid pressure). A total of 41 data points that show coverage across a selection of attributes (API, depth, FVF, pressure and temperature) have been selected for further analyses.

Both basins have a good spatial distribution of data, and despite both being classed as 'mature', they remain prolific with recent discoveries being made. As such, they make strong candidates for 'big data' mining and machine learning.

The 6391 data points from *Basin X*, and 41 data points from *Basin Y* with good coverage have been used to identify dependencies between attributes. The remaining 4165 data points in *Basin X* are incomplete, and do not have coverage across all the attributes. A primary aim of this research is to identify possible dependencies within multivariant reservoir datasets. Such dependencies would allow data coverage to be increased through the prediction of missing values using machine learning techniques. To maximise the impact of increasing data coverage, attributes that are required as input to the hydrocarbons initially in place (HCIIP) calculation (Figure 2) were high-graded for analyses. STOIIP (stock-tank oil initially in place) and GIIP (gas initially in place) values are not currently available for all data in the basins (Figure 1). Increasing the coverage of HCIIP estimations across the basins will support more accurate reserves assessment and aid decision making in exploration and portfolio management. Seven attributes are required to predict HCIIP: area, gross thickness, net thickness, porosity, oil saturation and the formation volume factor (FVF) (Figure 2). Six additional attributes (API, depth, permeability, pressure, temperature, viscosity) were included to capture potential dependency to the STOIIP input attributes (Figure 2). These twelve attributes have been chosen to test our presented data mining workflow.

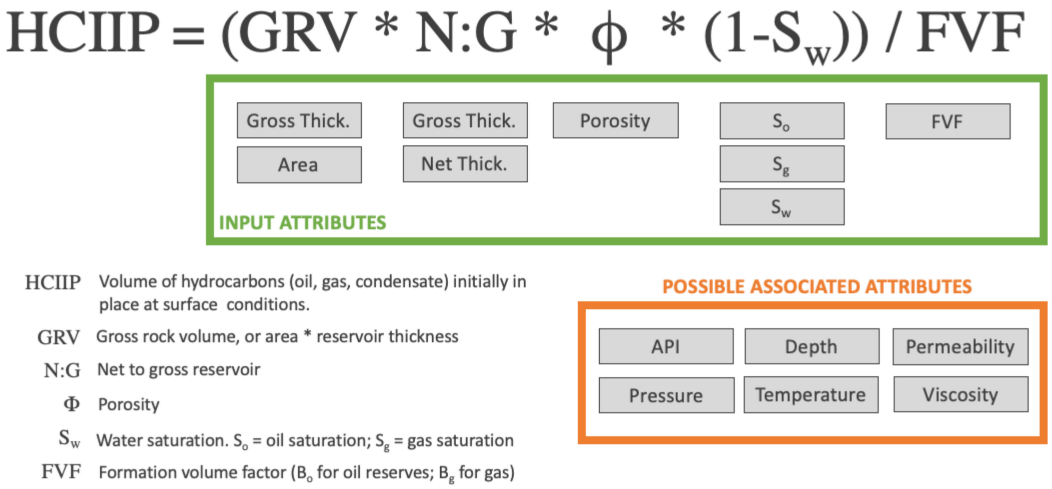

HCIIP = (GRV * N:G * ϕ * (1-S_w)) / FVF
Gross Thick.
Area
Gross Thick.
Net Thick.
Porosity
S_o
S_g
S_w
FVF
INPUT ATTRIBUTES
HCIIP Volume of hydrocarbons (oil, gas, condensate) initially in place at surface conditions.
GRV Gross rock volume, or area * reservoir thickness
N:G Net to gross reservoir
Φ Porosity
S_w Water saturation. S_o = oil saturation; S_g = gas saturation
FVF Formation volume factor (B_o for oil reserves; B_g for gas)
POSSIBLE ASSOCIATED ATTRIBUTES
API
Depth
Permeability
Pressure
Temperature
Viscosity

Figure 2. The Hydrocarbons Initially In Place (HCIIP) calculation and relevant attributes within the datasets. Primary attributes directly feed into the HCIIP calculation. Six additional attributes are identified as likely to show some dependency to the HCIIP attributes.

2.2. Workflow and Methods

The data-driven AI workflow presented herein comprises the following steps:

1. *Exploratory data analysis*. This first step (Figure 3) constrains the available attributes, their coverage, and variability. Data QC can be done using basic exploratory statistical visualisation tools, for example: histograms; cross-plots; and low-dimensional linear projection. Histograms are used to visualise the data distribution. Cross-plotting each attribute against depth highlights outlier data points and general trends with depth. Finally, linear projection plots are used to assess multivariate data clustering and outliers [25,26]. A circular placement was used which plots attributes on independent

vectors, and allows the visualisation of multidimensional data in two-dimensional space [26]. This allows attributes to be assessed independently thereby assuming no attribute has a higher importance for clustering (as in principle component analysis). The data visualisation was completed using Orange Data Mining software [27].

2. *Detection of outliers.* Clustering of multivariate data allows outliers and erroneous data to be separated from bulk data prior to further analysis. It is important that outliers are treated separately in the prediction workflow so they do not bias prediction on bulk populations and to avoid imbalance problems [28,29]. Data separated from the bulk population may represent legitimate outliers or small populations behaving differently for a geologically plausible reason, e.g., a giant field or reservoirs with exceptional characteristics due to overpressure or historic subaerial weathering. Hierarchical clustering is used in this study to separate outliers and distinctive groups of geological populations. This agglomerative cluster analysis method groups the data by their proximity in multidimensional space according to the chosen distance metric [30–32]. This can efficiently separate outliers in addition to identifying distinct populations of data. Hierarchical clustering was completed using Orange Data Mining software [27].
3. *Grouping similar reservoir and fluid characterisation features*. This step identifies groups of reservoir and fluid variables that exhibit similar trends within the data using self-organising maps (SOM) [33]. This unsupervised learning tool groups attributes showing similar dependencies in a reduced space—a 2D lattice. SOM uses competitive learning to assign the input data to neurons on a lattice as arranged by their similarity. The distances between the data on the 2D lattice are visualised by heat maps, where cool colours indicate small distances between data points (clustering), and warm colours represent large distances between data. Here, we use SOM lattice data projection to generate a heat map for each attribute that depicts mutual relationships between the trends in the data, therefore allowing geological attributes to be grouped based on the trends they exhibit. In this data-driven workflow, this step ensures only those physically meaningful trends are brought forward for predictions. SOM was completed using the software 'ML Office' [30].
4. *Feature selection.* This step in the workflow explores input/output predictor structure in order to select the most relevant combination of inputs for predictions. K nearest neighbour (KNN) is a straightforward deterministic estimation method based on the similarity of data which is in close proximity in multidimensional space [34]. The regression estimates are computed based on the optimum number of neighbouring data points. The key parameter is the number of K neighbouring data points to be optimised for the most accurate prediction. We demonstrate that this basic method, which uses a single parameter to tune, suits the purpose of finding the combination of those input features that are most relevant to the target variable. Cross-validation is commonly used to find the optimal K value at the minimum of the cross-validation error curve [30] (Figure 4). The shape of the error curve can reveal the relevance of the input dimensions to the output (target) variable. Where the error curve has a declining shape with increasing number of neighbours without a distinctive minimum (Figure 4a,b), the target variable does not depend on the distance between the data in the input space and the combination of inputs is, therefore, not relevant to the target variable. An error curve that features a distinct minimum corresponds to a KNN estimate for a target variable combination that can be spatially correlated within the input space (Figure 4c,d). This not just depicts the optimal number of neighbours for a KNN estimate but more importantly confirms that the input attributes are relevant input features that define the variation of the target variable values in the input space [30]. Feature selection using KNN tells us which combination of the input attributes are most relevant to make predictions. KNN cross-validation was completed using the software 'ML Office' [30].
5. *Prediction of reservoir properties to increase data coverage.* In this study, we test the use of Random Forest (RF) to make predictions and fill gaps in attribute values based

on the selected relevant features. Random Forest regression is a tree-based learning algorithm proposed by Breiman [35], and has been successfully used for classification of geological data in previous studies [36]. The regression returns the average prediction of individual trees within an ensemble, or forest [37] (Figure 5). A decision tree makes a decision after following a classification from the tree's root to its leaf. The decision path includes nodes, branches, and leaf nodes. Each node represents a feature, each branch describes a decision, and each leaf depicts a qualitative outcome. RF constitutes the combinations of features that make predictions based on the probability assigned to branches. The probabilities are updated whilst data are propagated through the RF in a supervised learning fashion. The RF regression prediction model was trained and tuned using training and validation subsets. A blind test set prediction was computed to demonstrate the overall performance of the predictions. Model generation and testing was completed using Orange Data Mining software [27].

6. *Prediction confidence*. Prediction confidence was assessed based on multiple prediction models with different inputs and hyperparameter tuning. Test error was used as a measure of confidence. Prediction confidence analyses were completed using Orange Data Mining software [27].

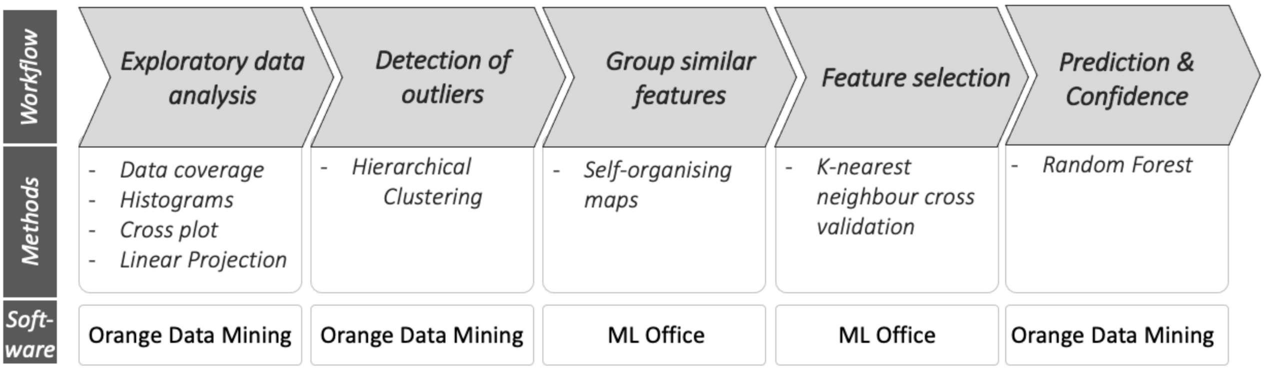


Figure 3. Workflow, methods and software used in this study. Software used included Orange Data Mining [27] and Machine Learning (ML) Office [30].

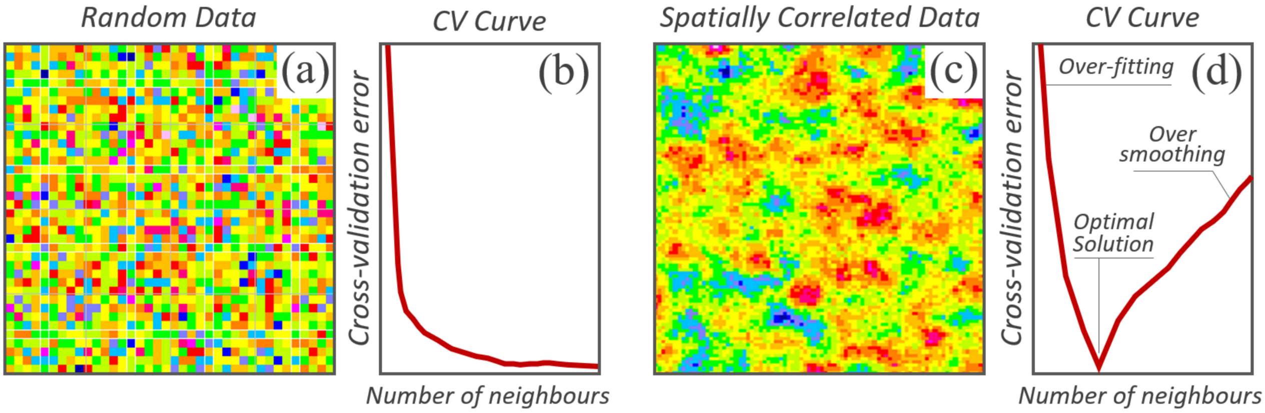


Figure 4. Feature selection with KNN: (**a**) uncorrelated (random) data pattern; (**b**) associated cross-validation error curve steadily declining for uncorrelated data; (**c**) spatially correlated data; (**d**) cross-validation error featuring a distinct minimum for correlated data (after [30]).

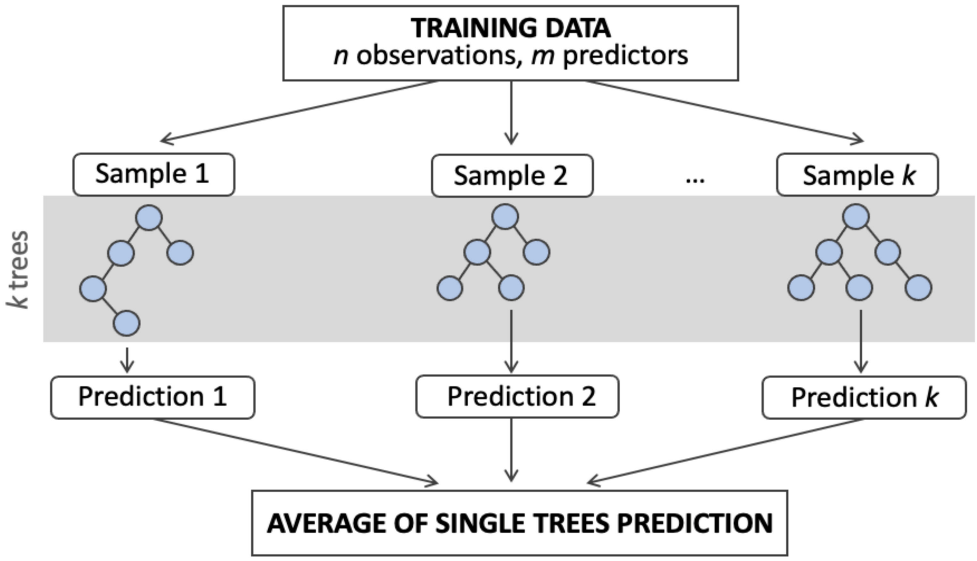


Figure 5. Random Forest prediction structure, after [38].

3. Results

3.1. Exploratory Data Analysis and Detection of Outliers

Exploratory data analysis uses several data visualisation methods to support a quick evaluation of data distribution statistically across multiple attributes. This approach checks for outliers in the data and provides a quality check for appropriate ranges of variables. Histograms of the input attributes in *Basin X* (Figure 6) show close to normal or slightly skewed distributions for API, depth, porosity and saturation. Other attributes show a heavily skewed data distribution with possible outliers, for example, area, net thickness, gross thickness and viscosity (Figure 6). The porosity, temperature and saturation data distributions indicate that the dataset used a mixture of integer and decimal data, and therefore does not show a smooth normal distribution (Figure 6).

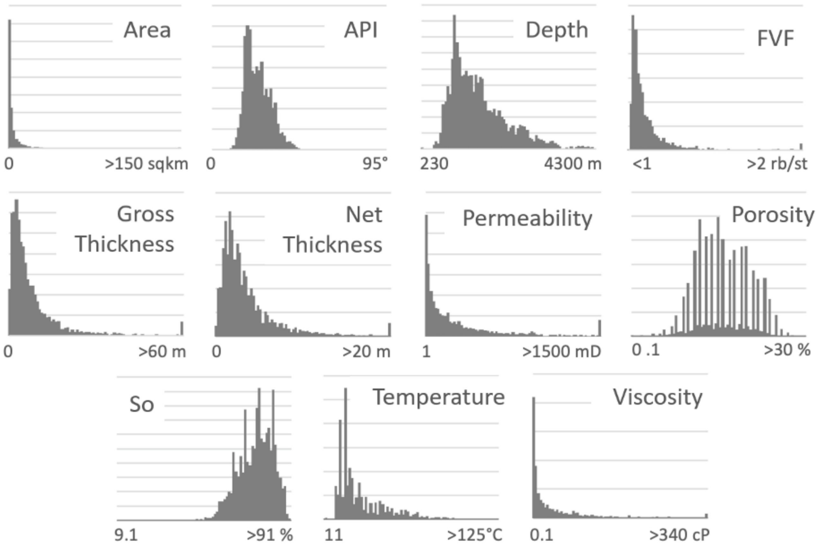


Figure 6. Histograms show attribute data distributions for *Basin X*.

Data from both basins have also been visualised in two dimensions to assess for outliers and general trends with depth (Figures 7 and 8). A number of attributes from *Basin X* (Figure 7) demonstrate some association with depth including API, FVF and Temperature. Despite a large scatter in the data, a decrease in porosity with depth can be detected. Other attributes show peaks at specific depths, with normal distribution (area, gross thickness) or bimodal distribution (permeability, viscosity). The data from *Basin Y* (Figure 8) show similar trends across the attributes. It is clear that a number of attributes show outliers, and different levels of scatter or noise within the data.

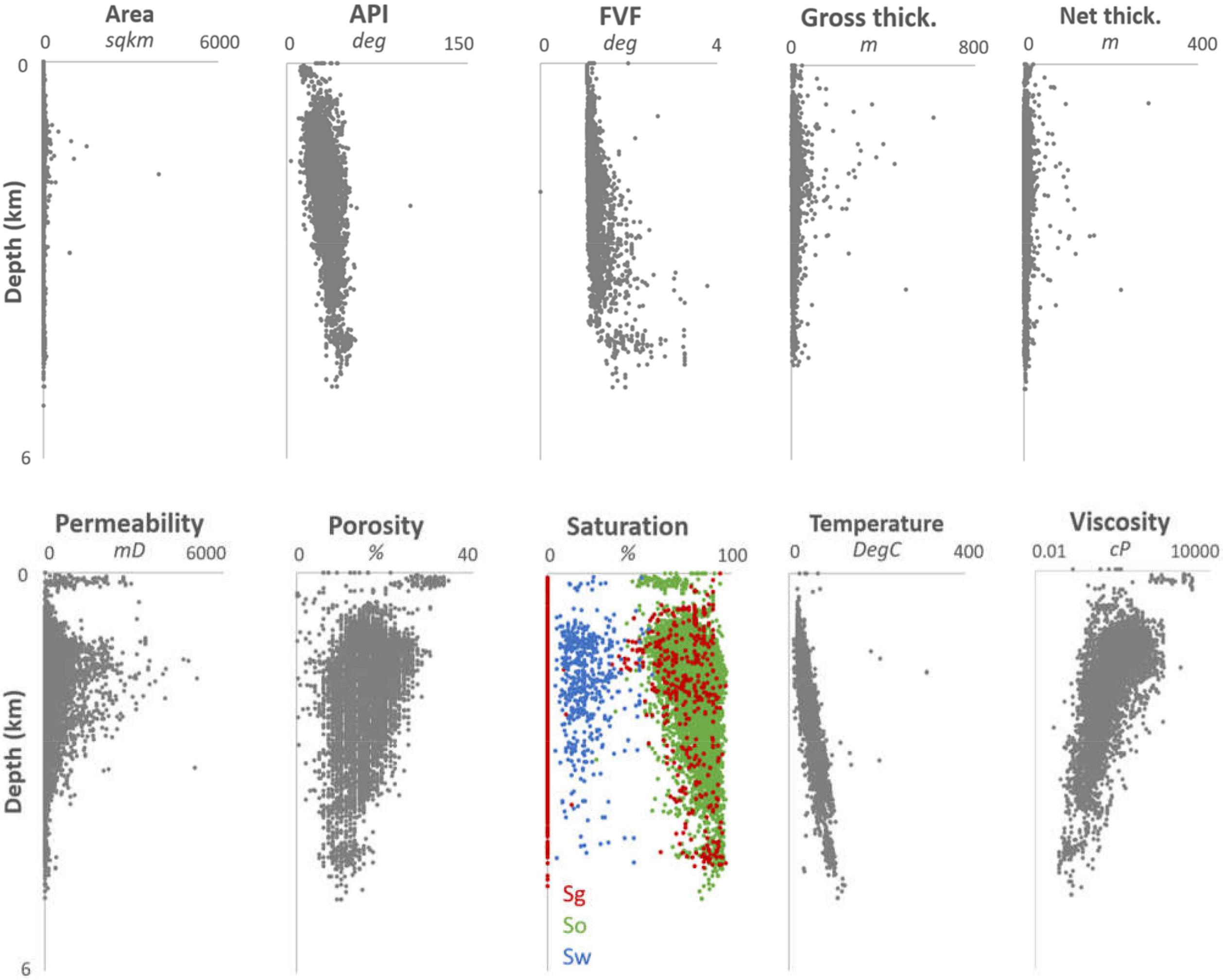


Figure 7. Cross-plots for various attributes against depth for *Basin X*.

The *Basin X* data have also been visualised in a multivariant linear projection to assess the relative separation of the outliers (Figure 9). Linear projection is a useful tool for the visualisation of multivariant data in two dimensions [25,26]. It can be used to visually assess clusters within the data, as well as identify outliers. Here, we use a circular placement which plots attributes on independent vectors [26]. This allows attributes to be assessed independently and assumes no attribute has a higher importance for clustering (as in principle component analysis). The resulting plot shows that area and viscosity have significant outliers, with likely additional anomalous data points seen in the FVF values (Figure 9). There are indications of multiple data populations, however, the majority of data tend to group into a distinct cluster of values of similar range.

It is clear from the initial data distribution visualisation (Figures 6–9) that the dataset contains a number of outliers across multiple attributes. These may be erroneous data points resulting from poor measurements or unit error, or they could be geologically valid. In the case of *Basin X*, it is confirmed that there are a small number (<5) of giant fields that do not conform to the general trends seen in the basin. Other values could also be geologically valid, for example, exceptionally high values for viscosity which could be in error, or due to local biodegradation.

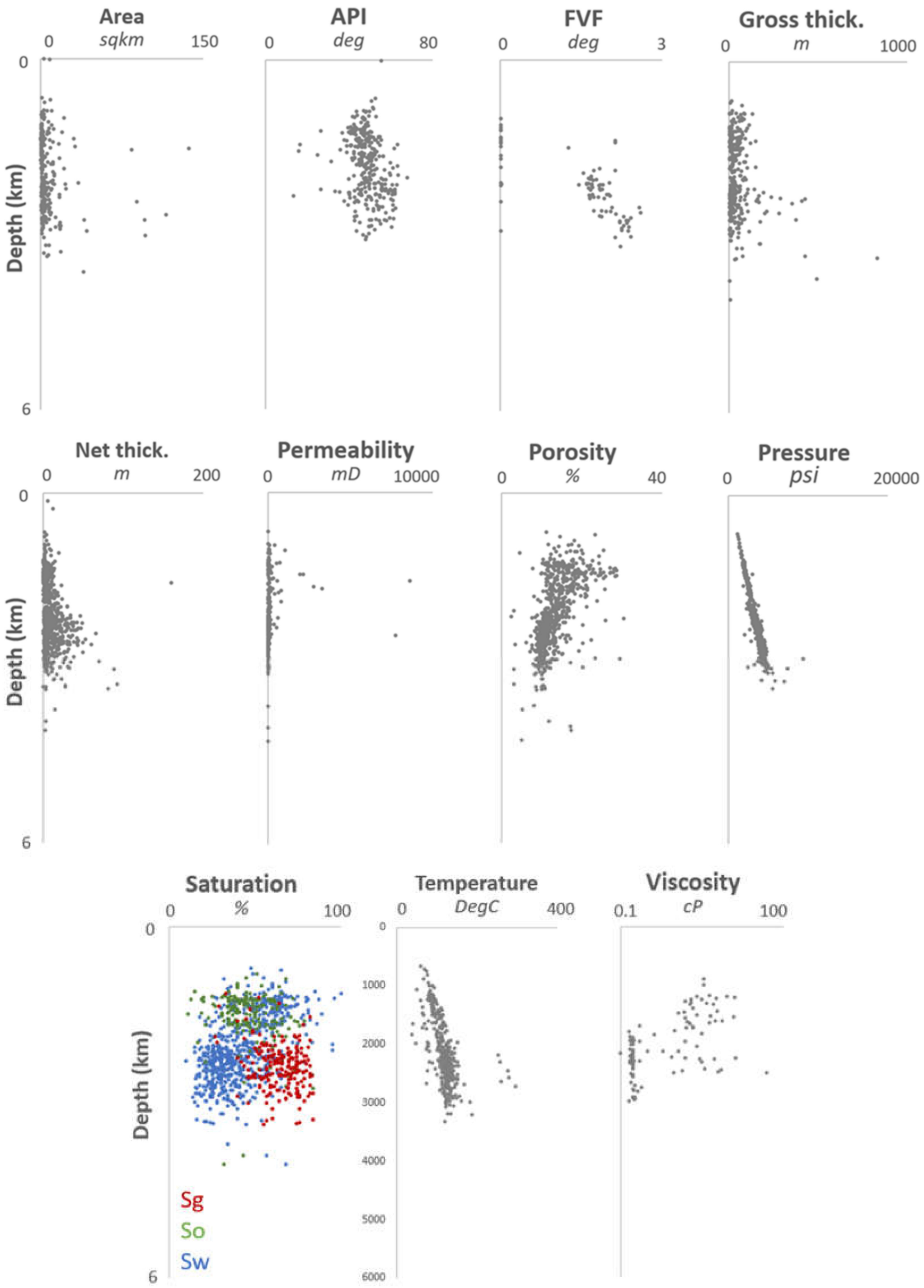


Figure 8. Cross-plots for various attributes against depth for *Basin Y*.

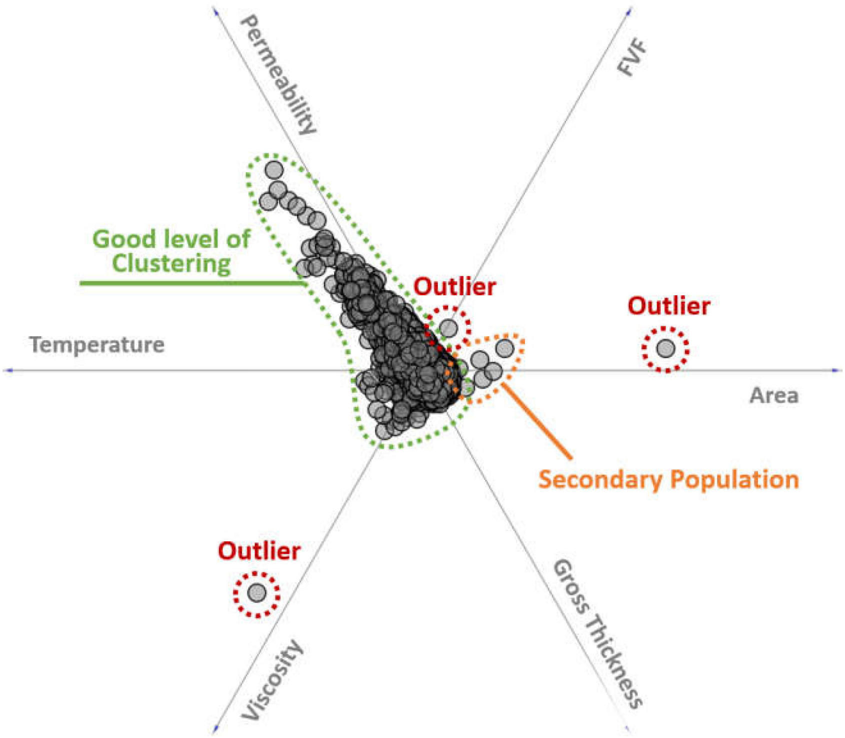


Figure 9. Normalised multivariant linear projection of attributes: area; FVF; gross thickness; permeability; temperature; and viscosity with outliers and possible secondary populations highlighted. Plot from Orange Data Mining software [27].

Despite the geological validity of some outliers in the dataset, all must be removed to ensure the identification of dependencies, allow predictability within the data, and avoid data imbalance problems [28,29]. To do this, hierarchical clustering methods were used [30–32]. Six clusters were identified in 11-dimensional space. Figure 10b indicates that the clustering is effective at separating outliers and secondary data populations from the dataset. Clusters 1–4 were rejected from the study as outliers, with cluster 5 taken forward for regression analysis (removing 28 outlier values) (Figure 10a).

3.2. Grouping of Similar Features: SOM

Cluster 5 (C5) data were analysed using self-organising maps (SOM). This unsupervised learning technique can be used to assess similarities in data distribution across multiple attributes, and identify the optimum number of nearest neighbours for machine learning. All eleven attributes were analysed for clustering using SOM. The resulting U-matrix map for the 11 inputs demonstrate the likelihood of distinct clustering in this multivariant data (Figure 11). The U-matrix heat map is not random, therefore, there are some relationships and dependencies within the data. This indicates that supervised learning in the form of KNN may be viable to assess dependencies across the attributes.

Individual input layers to the U-matrix reveal that some attributes show very similar distributions and input layers can be grouped based on their character (Figure 12). Most notably, API, depth, FVF and temperature show close similarities. Each of these variables shows a clustering to the bottom right of the input layer, with increasing separation of data points in the top left. Oil Saturation, S_o, may also follow this general trend, but demonstrates added complexity and less clustering. Viscosity data plot in the opposite trend.

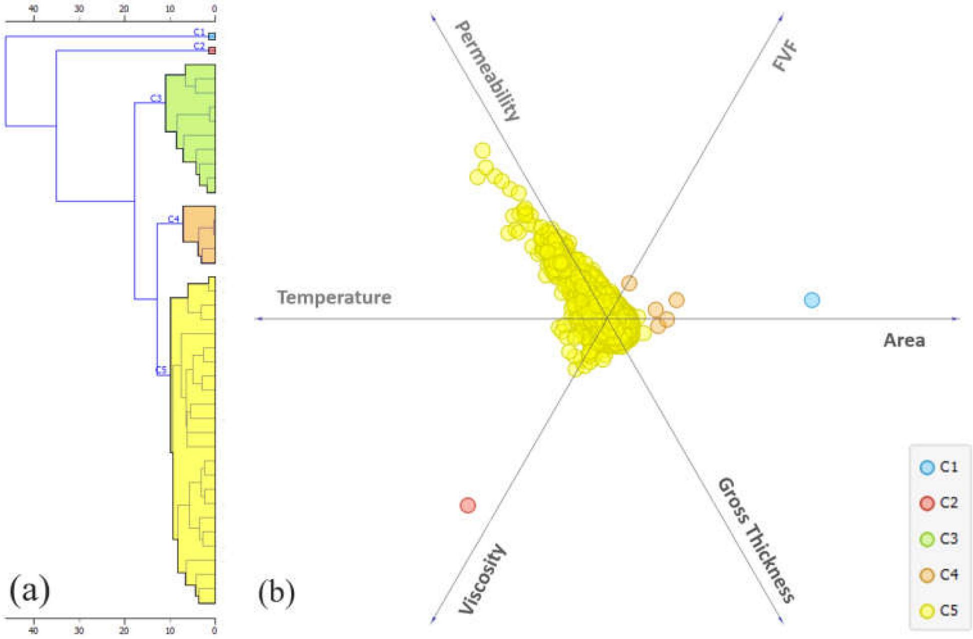


Figure 10. (**a**) Hierarchical clustering of *Basin X* data into 5 distinct clusters in 11-dimensional space. (**b**) Multivariant linear projection of attributes area, FVF, gross thickness, permeability, temperature and viscosity. Colours indicate Clusters C1–C5 and demonstrate that outliers are captured well by hierarchical clustering. Plot from Orange Data Mining software [27].

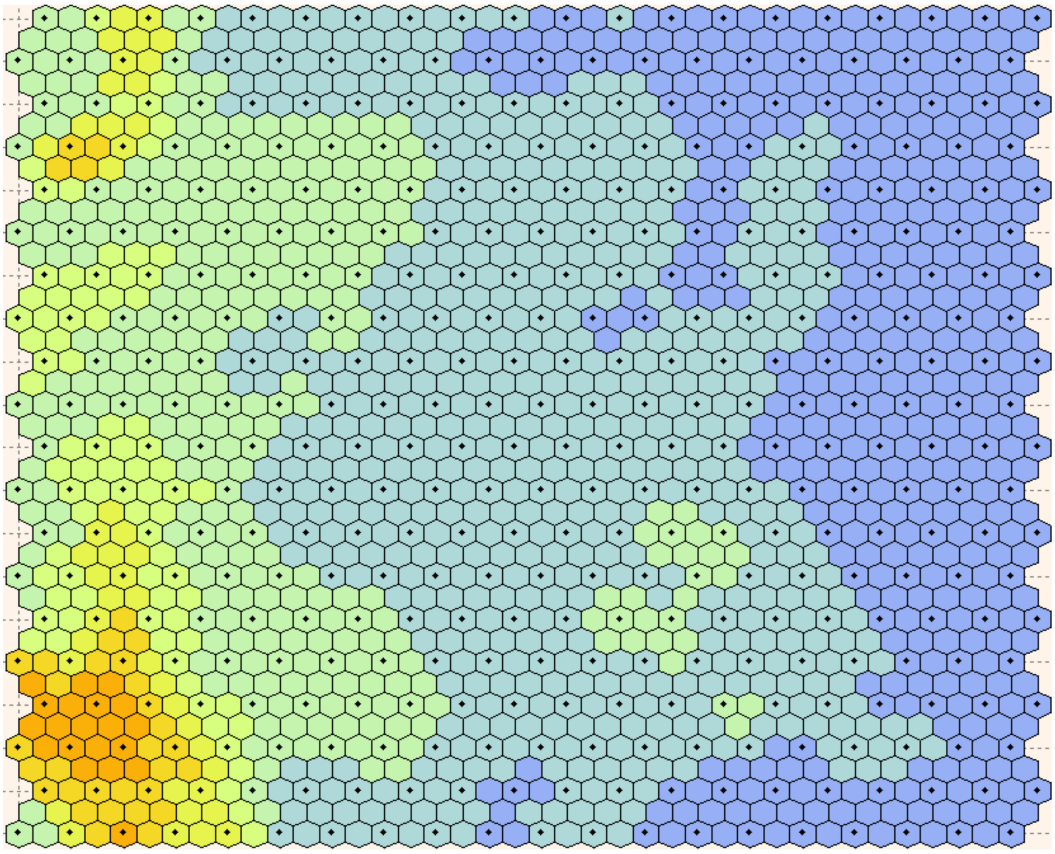


Figure 11. U-matrix for the 11 inputs (shown in Figure 12). Cool colours indicate small distances between data points (clustering), warm colours represent large distances between data in multidimensional space. SOM maps generated in ML Office [30].

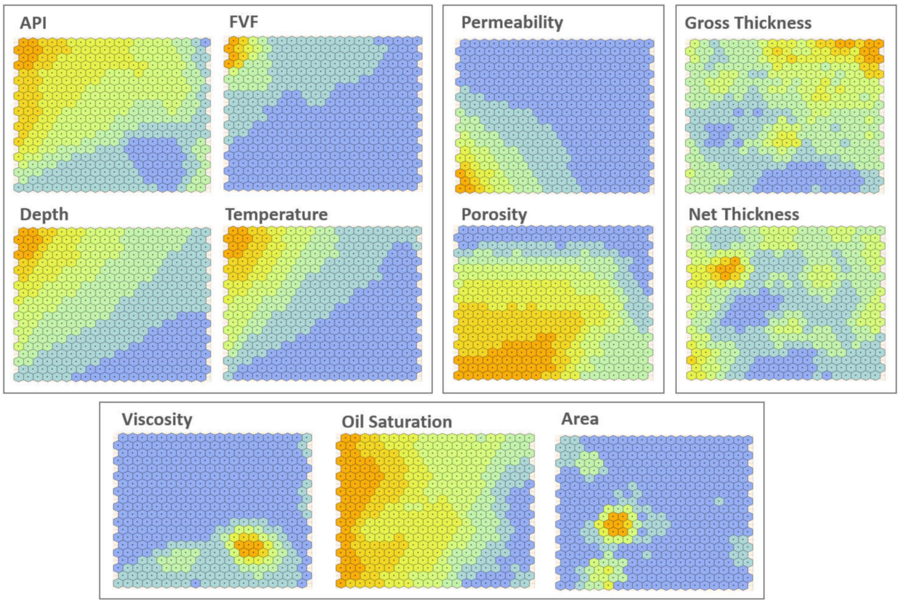


Figure 12. Heat maps indicate the individual input layers that generated the final SOM U-matrix (Figure 11). Input layers have been grouped based on data trends. In all maps, cold colours indicate small distances between data points and warm colours indicate large distances between data points.

Porosity and permeability show a distinct linear trend, clustering in the top and right of the of the input layer. There are similarities in the trends of porosity and permeability input data, although the porosity data appear to show more discrete clustering. Other variables show more complex clustering of data. Net and gross thickness show similarities, and both show multiple clusters in the dataset. Area data appears to be made up of multiple clusters quite unique from the other data distributions (Figure 12).

Due to a lack of pressure data in the *Basin X* dataset, SOMs were constructed for *Basin Y* for the three attributes depth, temperature and pressure (Figure 13). API, FVF and viscosity have been excluded due to insufficient data coverage hampering the identification of statistically-meaningful dependencies. The resulting U-matrix heat map is not random (Figure 13), indicating that there are dependencies between the variables present in the dataset. Nevertheless, it is challenging to identify the number of possible clusters based on visual inspection of the U-matrix map. Heat maps of the three input layers show similar distributions, with clustering occurring in the top right of the input layer, with increasing separation of data points towards the bottom left of the input heat map (Figure 13).

The SOM analysis of *Basin X* suggests a strong association between API, depth, FVF, and temperature based on similar trends mapped on the SOM input layers (Figure 12). Therefore, these four attributes were chosen to review feature dependencies and to select the combination of input attributes that optimise the prediction accuracy. The aim of this step is to identify which combinations of input attributes are most relevant to predict others. We used a simple KNN cross-validation method to demonstrate how the relevancy of the input variables changes when more attributes are used for prediction. We conducted experiments using one and two input attributes to demonstrate the change in relevancy with the number of inputs. We also tested the relevance of viscosity, which appears to show an inverse trend with depth compared to API, FVF, and temperature (Figure 12) and conducted a final experiment using all four input attributes.

3.3. Feature Selection: KNN

In the first experiment, pairs of attributes were checked to see if one could be used to predict the other in a simple linear or non-linear regression in two-dimensional space (Figure 14). Cross-validation (CV) error curves demonstrate a systematic decline with no strong minima. This means no single attribute is relevant to another in two-dimensional space.

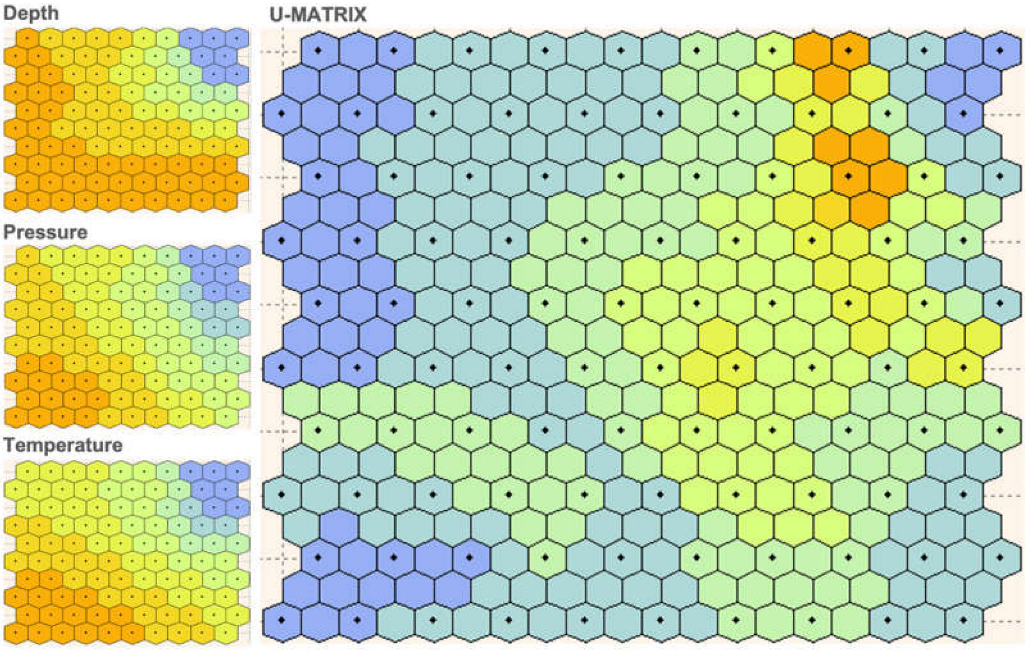


Figure 13. Input layers for depth, pressure and temperature, and the resulting U-matrix for all attributes for *Basin Y*. Cool colours indicate small distances between data points (clustering), warm colours represent large distances between data in multidimensional space.

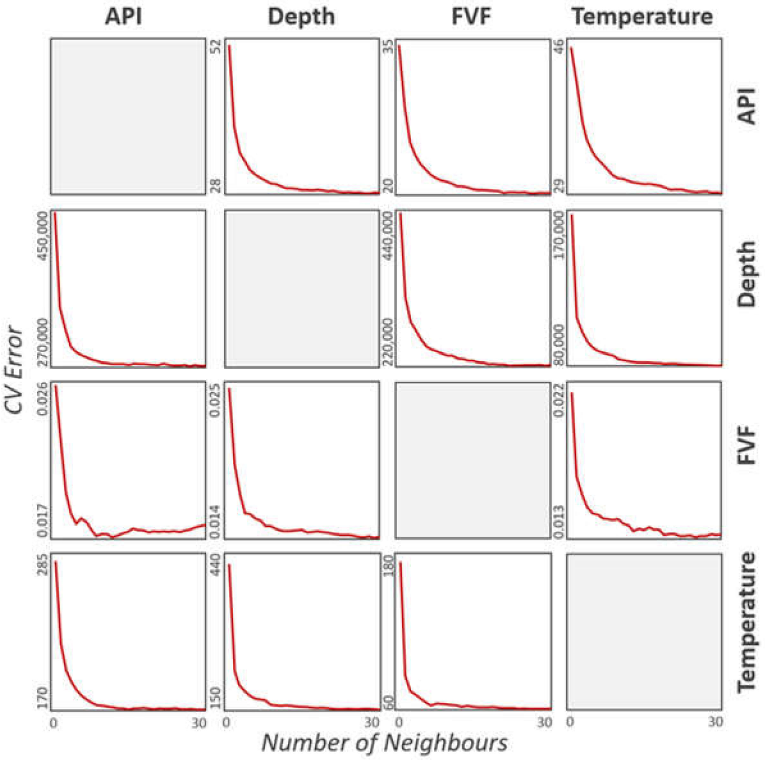


Figure 14. KNN cross-validation (CV) error curves for each variable (API, depth, FVF, temperature) in 2-dimensional space. The input feature is listed on the horizontal axis, and the target variable listed on the vertical. Where the error curve (red) has a declining shape with increasing number of neighbours (*x*-axis), the target variable does not depend on the distance between the input data and is therefore not relevant to input features.

In the second experiment, we used combinations of any two attributes to predict a third (Figure 15). Some of the combinations demonstrate noticeable minima of the CV error curves, indicating those pairs of inputs can be deemed relevant for predicting the target attribute. Key observations are:

- *API:* The CV error minimum corresponds to K = 5 neighbours with FVF and temperature as the inputs (Figure 15a). An increasing CV error with increasing neighbours indicates a level of predictability based on the clustering of these data. Other combinations—FVF and depth, depth and temperature, show little relevancy to predict API with no apparent minima on the error curve (Figure 15a). This suggests that depth has little relevance to API values in this reservoir dataset.
- *Depth*: API and FVF appear to be the most relevant combination of inputs to predict depth. API and temperature show a very weak (but still detectable) minima (Figure 15b).
- *FVF*: Any pairs of attributes appear to be relevant to predict FVF. K = 2 neighbours give the minimum error with API and temperature attributes as inputs. Depth and temperature have the optimal K = 4, and API and depth input attributes give an optimal K = 8 (Figure 15c).
- *Temperature:* Clustering of temperature data is seen with respect to API and FVF, with little to no relationship identified with any other combination (Figure 15d).

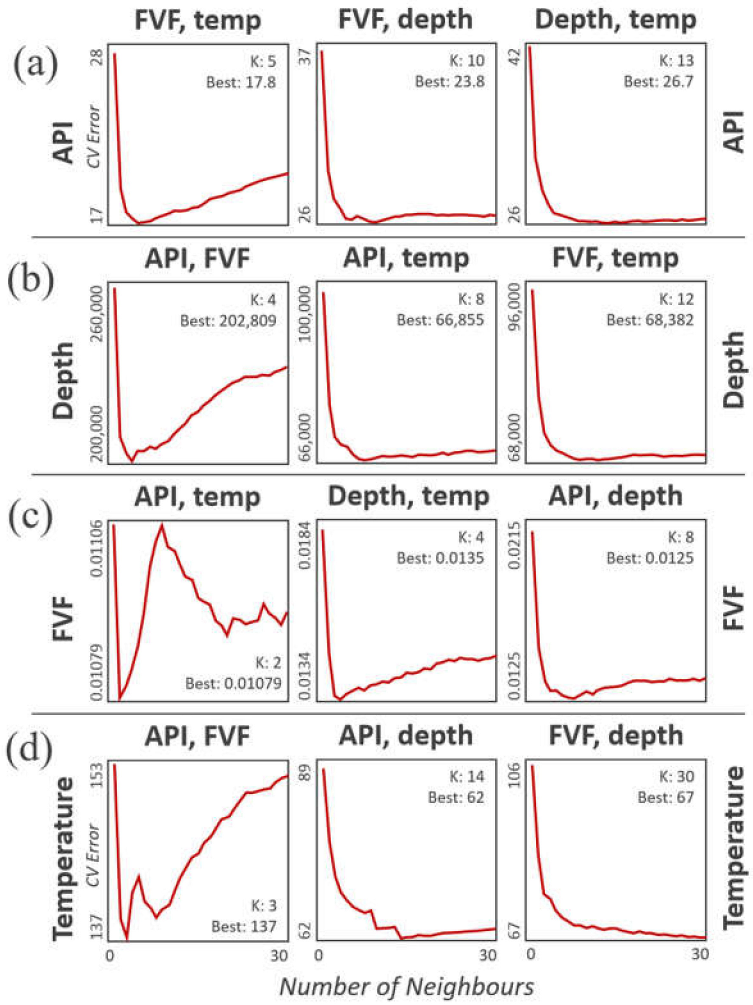


Figure 15. KNN cross-validation (CV) error curves for each variable in three-dimensional space. The input features are listed on the horizontal axis, and the target variable listed on the vertical where: (**a**) API; (**b**) depth; (**c**) FVF; (**d**) temperature. Where the error curve (red) has a declining shape with increasing number of neighbours (x-axis), the target variable does not depend on the distance between the input data and is therefore not relevant to input features. An error curve that features a distinct minimum corresponds to a KNN estimate for data that can be spatially correlated.

Finally, we used four input attributes to analyse the relevancy. Viscosity was added to the pool of attributes as it showed the inverse trend relative to API, depth, FVF and temperature (Figure 12). An associated, but inverse trend indicates that it also has a dependency to the other four attributes. The CV error curve demonstrates strong minima for all combinations of the 4 variables to predict the 5th (Figure 16). The optimal number of neighbours varies from K = 4 to predict FVF and depth, K = 5 for API prediction, K = 10 for viscosity prediction and K = 19 to predict the temperature. Temperature shows the weakest minimum in the CV curve, indicating poor relevancy in four-dimensional space compared to using just API and FVF as input features (Figure 15).

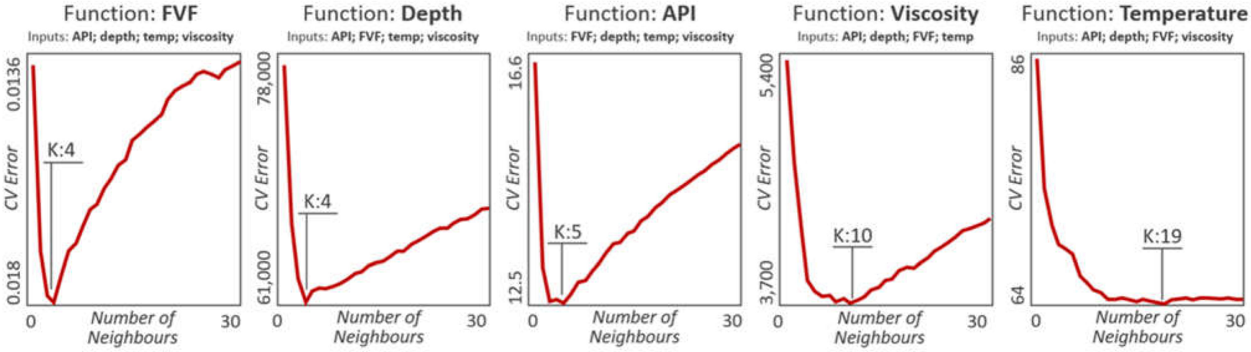


Figure 16. KNN cross-validation (CV) error curves for each variable (API, depth, FVF, temperature, viscosity) in five-dimensional space.

3.4. Prediction and Confidence

We have evaluated the predictability of FVF based on the inputs API, depth, temperature and viscosity. FVF has been chosen as a test case as it is an important input to the HCIIP equation (Figure 2) and shows a strong level of potential dependency to the four input variables in the data as indicated in the KNN cross-validation curve (Figure 16). Outliers were detected and removed with hierarchical clustering to avoid the impact of abnormal values that may distort the predicted dependencies. Ten clusters were used to remove 44 outliers. Clusters C10, C7 and C5 were selected for modelling (Figure 17a). The 6348 data points were split into training (75%) and test (25%) datasets. Prediction of missing attribute values was computed using a supervised learning algorithm: Random Forest (RF) [32]. The model was built with 20 decision trees and a minimum number of splits limited to 5 to control growth.

When the RF-predicted FVF values are plotted against depth (Figure 17b), the distribution is remarkably similar to the recorded FVF (Figure 17a). However, when random forest data points are coloured based on their original FVF, some errors are clearly seen, particularly for high FVF values (identified in yellow in Figure 17b). Model evaluation (Table 1) gives a mean squared error (MSE) of 0.004, and a R-squared value of 0.861. R-squared is used to determine how well the model fits the data, with 1.0 indicating a perfect fit. In this case, the model R-squared evaluation indicates a good level of predictability at 0.861, or 86.1%.

Table 1. Prediction error and effectiveness for various input scenarios. Three models were run: one on the complete dataset (including outliers), one for clusters C10, C7 and C5, and one for C10 only. MSE = Mean Square Error; R2 = R-squared, or the coefficient of determination.

Input Clusters	MSE	R2
Clusters C1–C10	0.004	0.858
C10, C7, C5	0.004	0.861
C10 only	0.002	0.828

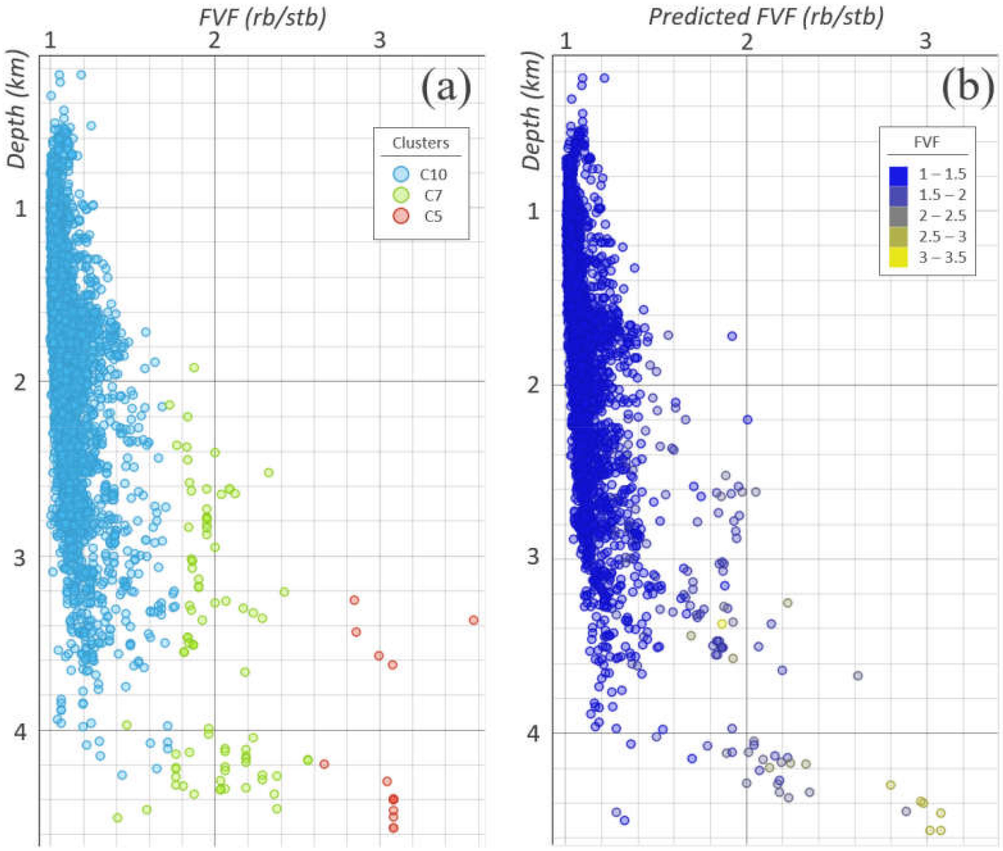


Figure 17. Results from random forest learning. (**a**) Recorded FVF values plotted against depth. Hierarchical clustering was used to remove outliers. Three clusters (C10, C7 and C5) were used for model training. (**b**) FVF predictions (based on API, depth, temperature and viscosity data) plotted against depth. Data points are coloured based on their actual recorded FVF value. Some regions of incorrectly predicated values are identified, particularly at greater depths. Cross-plots modified from Orange Data Mining software [27].

The model was also run for the full dataset (Clusters C1–C10) in order to assess the impact of outliers in the data. Model evaluation gives a MSE of 0.004 and R-squared value of 0.858 (Table 1). This minor reduction in the accuracy of the predictions indicates that it is beneficial to complete data cleaning prior to modelling. Finally, the model was run using cluster C10 only—removing the high FVF populations (Figure 17a). This also marginally reduces the level of predictability (R-square value of 0.828, Table 1). Importantly, the removal of Clusters C5 and C7 could lead to FVF values being under-estimated in deeper wells. It is, therefore, necessary to assess the geological validity of clusters in the data, and balance model accuracy while also ensuring the model is fit for purpose.

Finally, the model was trained using different input attributes to assess model effectiveness when using 2-, 3- and 4-input attributes. This is a likely scenario when incomplete datasets have partial data coverage, and fewer than optimum number of input attributes (as is the case for *Basin X* which is lacking pressure data). Table 2 demonstrates that using all available attributes increases the accuracy of the model, however, there is a fair-to-good level of predictability where only 3 or 2 input variables are used. Of note, error is comparable when FVF is predicted using two variables (temperature and viscosity: MSE = 0.004, R2 = 0.845), or all four variables (API, depth, temperature and viscosity: MSE = 0.004, R2 = 0.861). These results demonstrate that despite all the attributes being relevant, not all are required for reasonable predictions to be made.

Table 2. Prediction error and effectiveness for the prediction of FVF across various input scenarios. Models were run to test model effectiveness for different input variable combinations to train the model. MSE = Mean Square Error; R2 = R-squared, or the coefficient of determination.

Variables	Input	MSE	R2
4	API, depth, temp, viscosity	0.004	0.861
3	API, depth, temp,	0.006	0.776
	API, depth, viscosity	0.005	0.811
	API, temp, viscosity	0.004	0.848
	Depth, temp, viscosity	0.004	0.853
2	API, depth	0.010	0.661
	API, temp	0.007	0.756
	API, viscosity	0.006	0.793
	Depth, temp	0.010	0.647
	Depth, viscosity	0.006	0.805
	Temp, viscosity	0.004	0.845

4. Discussion

4.1. Reservoir and Fluid Property Dependencies

Unsupervised Learning through SOM provides an initial assessment of trend similarities in large datasets. Individual input layers to the U-matrix reveal similarities in data distribution for a number of attributes (Figure 12). These similarities are particularly marked for the API, depth, FVF and temperature attributes, and viscosity which shows the inverse trend. We herein group these attributes as ‘reservoir fluid attributes’ since they relate to the hydrocarbon fluids within the reservoir.

Cross-plots confirm that attributes API, temperature and FVF systematically increase with depth (Figure 18), and viscosity decreases with depth (Figure 18). Key observations are: (i) a strong relationship between depth and temperature, with clear outliers; (ii) a general increase of API with depth; (iii) a subtle, nonlinear increase in FVF with depth; and (iv) a systematic decrease in viscosity with depth. Despite cross-plots showing clear trends (linear, polynomial and exponential), 2-dimensional cross-validation of KNN shows little evidence for clustering, and therefore, low levels of predictability (Figure 14). The use of multiple variants is therefore required to predict missing values.

Cross-validation of KNN in three- and five-dimensions (Figures 15 and 16) indicate that there is some level of predictability within the reservoir fluid attributes. To improve predictability, attributes may benefit from selective input where they show higher levels of predictability in multidimensional space. For example, when four attributes (API, depth, FVF, and viscosity) are used as inputs, temperature demonstrates a high k value, and no marked cross-validation error with increasing number of neighbours (Figure 16). This suggests that there is not a strong relevance between temperature and the other four attributes together. However, the CV curves in Figure 15 show a higher degree of predictability when dependencies to API and FVF are assessed only.

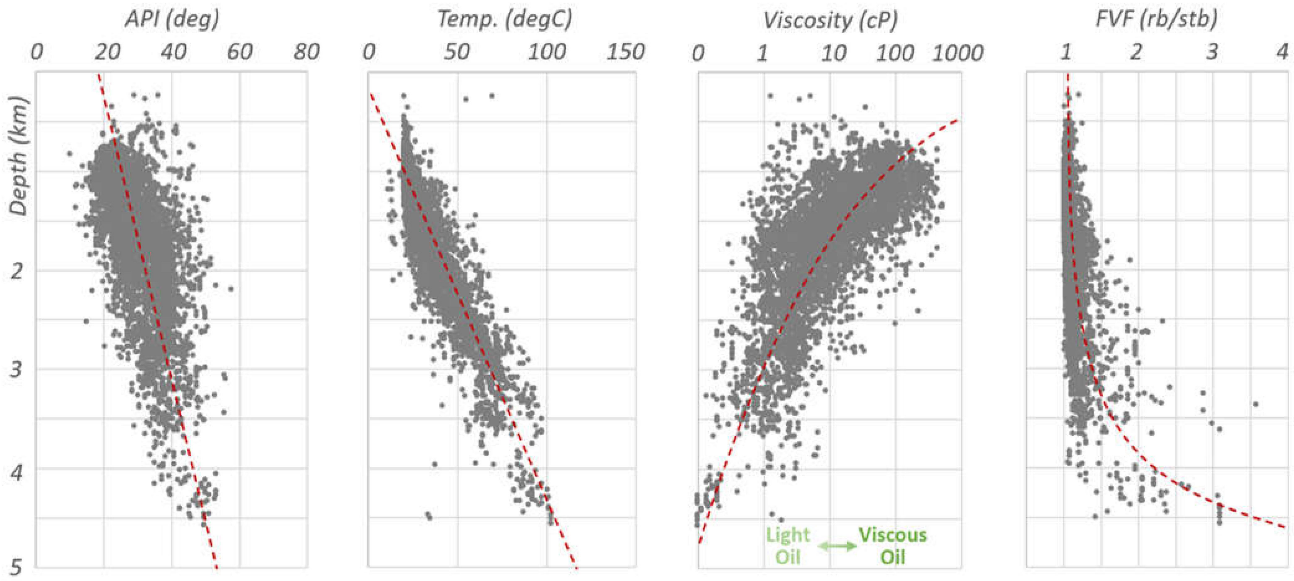


Figure 18. API, temperature, viscosity and FVF trends with depth in *Basin X*.

The SOM and cross-validation of KNN presented for *Basins X* and *Y* demonstrate a level of dependency between API, depth, FVF, pressure, temperature and viscosity. This is largely in agreement with known subsurface trends:

- *Temperature-Depth Relationship:* It is well documented that in the subsurface, temperature tends to increase with depth. The geothermal gradient is a measure of the change in temperature with depth and is closely related to the thermal conductivity of the rocks in the subsurface. Temperatures increase with depth due primarily to the decay of radioactive elements, such as potassium, thorium and uranium within minerals. Gradients can increase locally where magma emplacement occurs, or where crustal lithology is rich in radioactive elements (as in the case for many granites). Locally, the geothermal gradient may also be lower than expected where highly thermal conductive facies such as salt are present in the subsurface. Additionally, recorded temperatures may be distinct from the geothermal gradient due to local drilling effects (measuring artifacts) or human error. In the case of *Basins X* and *Y*, both show a linear relationship between depth and temperature, indicating that there are no unusual regions of heat flow (Figures 7, 8 and 18). This may not be the case in all basins, and care should be taken in volcanically active regions and salt basins. We note that *Basin Y* displays six values that are anomalously high but appear to increase with depth on a separate gradient (Figure 8). Further inspection of these values suggests that they represent measurements in degrees Fahrenheit (while the remaining dataset is in degrees Celsius). Such anomalies demonstrate the importance of step 1 of our workflow: exploratory data analysis, and its role in identifying incorrect or inconsistent units or data entry points.
- *Pressure-Depth Relationship:* No pore-fluid pressure data are available in the *Basin X* dataset; however, *Basin Y* data show a clear linear increase in pressure with depth. A linear increase in both hydrostatic and lithostatic pressure is expected with depth where fluids are in communication. Deviations from normal hydrostatic pressure can occur when fluids cannot escape and become over-pressured. Such outliers may limit the ability to accurately make predictions (regardless of their validity).
- *API-Depth Relationship*: Fluid gravity (API) is a measure of how heavy or light a hydrocarbon liquid is, and tends to increase with temperature, and therefore, depth. API values greater than 10 are those fluids lighter than H_2O. Therefore, any values in the dataset lower than 10 can be deemed erroneous, or measured in the water leg and provide no information on the hydrocarbon fluids within the reservoir. The higher the API gravity value, the lighter the hydrocarbon fluid. Therefore, fluids with higher API

values (gas, gas condensate) will accumulate in hydrocarbon traps above 'heavier' fluids (oils). Where traps are filled to spill, the later addition of gas may cause the displacement and remigration of oil to shallower reservoirs. Where underfilled, oil and gas columns can both be present in the trap. *Basin X* is a predominantly oil-prone basin and shows a linear trend of increasing API with depth. This is consistent with the assumption that source rock maturation increases with depth. The large degree of noise in the data could be accounted to different migration distances from the source rock. Abnormally low API values at shallower depths have been reported in other basins (e.g. in the North Sea [39]) as a result of biodegradation in lower temperature reservoirs. However, no significant modification to the API values due to biodegradation are seen in the *Basin X* dataset (Figure 18). *Basin X* is an oil-dominated basin, however, where a basin is mixed oil-gas, it is expected that API values will vary with depth. This is what is observed in the *Basin Y* API-depth cross-plot (Figure 8) which shows a much greater degree of scatter. This suggests that API would be more challenging to predict in a mixed oil-gas basin.

- *Viscosity-Depth Relationship:* Viscosity is a measure of the amount of resistance to flow the oil displays. Higher values indicate more resistance to flow. Viscosity is known to be closely associated with API, Temperature and Pressure [16]. When used with compositional data in the form of Watson's characterization factor [40], a clear inverse relationship between chemical composition, API and viscosity of oil can be seen. Viscosity is known to decrease with increasing temperature and pressure (up to the bubble point). In *Basin X*, the relevance of viscosity is confirmed by the KNN cross-validation curve against the other reservoir fluid attributes (Figure 16), and a good level of predictability is expected. The viscosity-depth plot shows a decreasing trend with depth (Figure 18). Note that viscosity plots on a log scale, which is an additional challenge when completing outlier detection and hierarchical clustering. A number of high cP values occur below ca. 250 m that can likely be attributed to heavy oils as a result or biodegradation.
- *FVF-Depth Relationship:* The Formation Volume Factor (FVF) is a key input to the HCIIP equation (Figure 2). For an oil accumulation, the oil formation factor corrects for the change in volume of oil at stock tank conditions compared to those under elevated pressure and temperature conditions in the reservoir. FVF is also closely linked to the level of gas saturation (GOR) [7]. *Basin X* shows increasing FVF values with depth on a nonlinear trend and increasing scatter with depth (Figure 18). This increasing variability of FVF with depth could be due to increasing gas solution, as is expected with the increasing API seen with depth.

The relevance of these five attributes has been shown through unsupervised (SOM) and supervised (cross-validation of KNN) learning. Review of these dependencies are geologically viable and are largely driven by increasing pressure and temperature conditions with depth (Figure 19). A number of valid explanations for outliers have been identified which may alter the data trends and care should be taken to understand the reason for outliers to determine if they should be removed from clustering analysis.

4.2. Machine Learning as a Predictive Tool for Reservoir Characterisation

A primary aim of this analysis was to determine if machine learning could be used to increase data coverage in large datasets, specifically those which are inputs to the HCIIP formula (Figure 2). We tested whether missing FVF values in the dataset could be estimated using API, depth, temperature and viscosity data. These attributes are a subset of the PVT (pressure-volume-temperature) properties that reservoir and production engineers require to characterise the reservoir fluids and plan for production. Our results show that FVF predictability is possible (Figure 17, Table 1) even where traditional correlation inputs are unavailable (e.g., GOR, pressure).

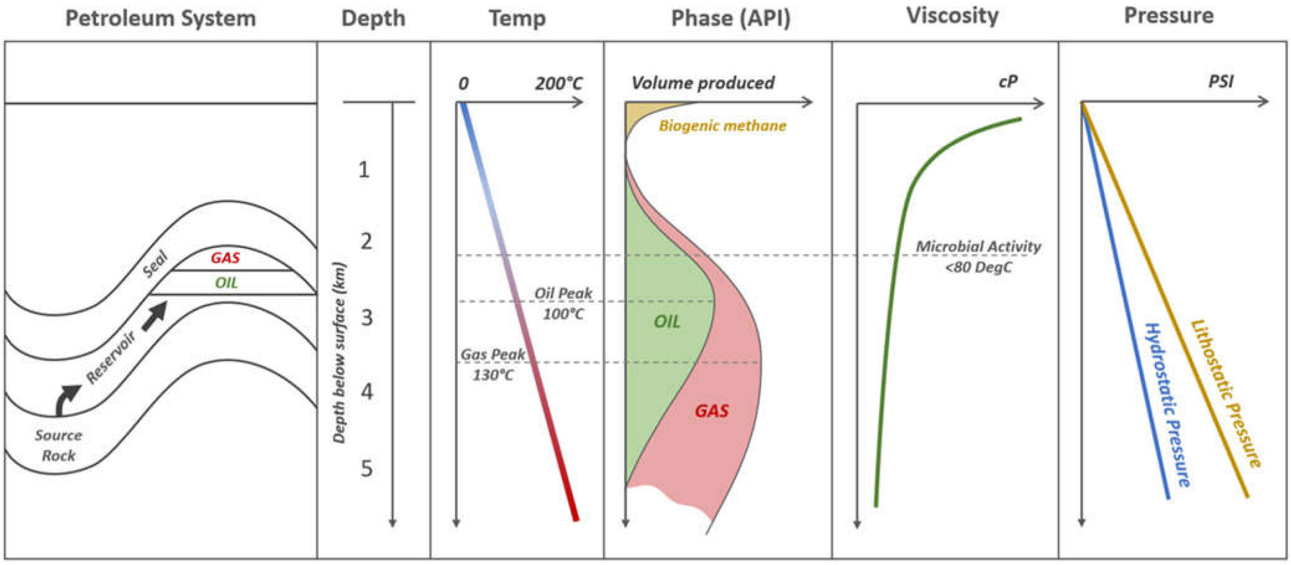


Figure 19. The approximate relationship between depth, temperature, API, viscosity and pressure and how these relate to the reservoir fluids. Adapted from [41].

Modelling error was tested using two, three and four input variables when predicting FVF (Table 2). Despite all the input attributes being relevant, Table 2 demonstrates that not all are required for effective predictions to be made. By using two or three selected inputs, data coverage can be increased without compromising model accuracy. In this case, just temperature and viscosity are required, with API and depth being relevant, but not essential for FVF predictability (Figure 15). In large datasets with partial input coverage, targeting the optimal predictor inputs will maximise the number of estimated data points added without a loss of prediction quality.

The example herein examined FVF, which forms one input to the HCIIP formula (Figure 2). Future research will assess predictability for remaining inputs: area, gross thickness, net thickness, porosity, and saturation. Initial results from unsupervised learning (Figure 12) indicate trends in the porosity, permeability and saturation data which may allow for a level of predictability. Area and thickness measurements may be more challenging to predict without additional geological information. Further work is also required to assess possible imbalance data artifacts for reservoir attributes, as we note that a number of attributes show a skewed distribution or inconstancies in data inputs (for example, integer vs. decimal input values (Figure 6)). Addressing possible data imbalance beyond hierarchical clustering techniques may increase predictability further.

Whilst large, incomplete datasets, such as those used in this study, are not the optimal source of detailed reservoir information to inform prospect- or field-specific information, estimation of missing values on basin-wide datasets provides valuable information to explorationists, allowing for basin screening and provide a better understanding of play- and basin-reserves. This can ultimately support data-driven decision making during the early stages of the E&P cycle. A strong over-riding geological understanding is required to assess machine learning results and assess the validity of data outliers and make geologically valid recommendations for subsurface risking. Combined data- and geoscience input is required to assess the nature of data, errors, interpretation and data dependencies and place these within a geological context.

4.3. *Future Applied Usage of Big Data in Subsurface Science*

This study has built a workflow for the assessment of reservoir data in two mature petroleum basins and tested if improvements in data coverage through the prediction of missing values is possible. In mature basins where exploration and production may have occurred for 50+ years, it is common to have erroneous or partial well data. This can be a factor of many reasons such as incomplete data suites being collected at the time of drilling due to the objectives of the well, cost, or technical constraints [9]. Data loss is also common through poor or inconsistent data management practices over the time since the well was drilled, or due to data being deleted by operators and service companies after their objectives have been met [42].

Data visualisation is a powerful tool to allow the geoscientist to assess the data quickly and efficiently and explore if data outliers are geologically valid or erroneous. Understanding data distribution may also identify unexpected areas of low (or high) coverage where data are absent or incorrect. We demonstrate that data visualisation in multi-dimensional space is beneficial for understanding data distribution, including outliers, and can inform inputs to stochastic simulation when calculating reserves. This data-driven approach could allow for a better quantification of subsurface uncertainty and economic risk in exploration and production [3]. Finally, increasing attribute coverage in large datasets can aid all aspects of the petroleum value chain from exploration, appraisal, development, production and enhanced recovery [3]. For example, increasing temperature coverage can reduce the risk in exploration. Temperature data are often absent or erroneous in legacy wells but are an important input to basin modelling to de-risk source rock generation and charge [41].

It should be noted that the applications of these datasets are not limited to the petroleum industry. The dependencies identified and discussed herein also have important relevance for emerging energies such as geothermal which relies on understanding temperature and pressure data. Subsurface storage of CO_2, Hydrogen and Methane will be integral to the energy transition. Utilising and repurposing datasets from mature hydrocarbon basins is a cost-effective way to gather important pressure, temperature and reservoir property data to assess and de-risk potential storage sites. There are many uses for 'big data' collected by the petroleum industry. However, for their effective use, careful data QC and exploratory data analysis with geoscience understanding are essential.

5. Conclusions

This study has demonstrated the use of a data mining workflow to constrain predictions of reservoir conditions and fluid properties using real-world datasets from two mature hydrocarbon basins. The sequential workflow addressed the following principle aims: (i) assess the adequacy of the data by streamlining the process of outlier detection, error detection and identifying natural artifacts in the data; (ii) identify dependencies within the data subject to the identified distinct groups of data clusters; and (iii) improve data coverage by estimating missing data values based on the identified data dependencies.

We note that outliers identified in the data are not always erroneous data values but may be geologically valid. This was seen in *Basin X* which contained a small number of giant fields significantly larger in area than the median field size. Other processes such as biodegradation generate secondary populations in multivariant datasets. It is therefore of key importance that the data scientist works closely with a subsurface specialist to ensure that data are not disregarded and disposed of incorrectly. Combined data- and geo-science input is required to assess the nature of data, errors, interpretation and data dependencies and place these within a geological context. Several demonstratable dependencies between attributes were identified in SOM and KNN analyses. Multivariate predictors used to estimate missing values require selection of the optimal combination of relevant attributes. Predictability within the data was demonstrated with an example of FVF: an important input to the HCIIP equation. Increasing data coverage in legacy datasets has important implications for reducing risk in hydrocarbon exploration, development and production.

Despite the many benefits of using data mining and machine learning techniques on legacy data, these techniques are still underutilised by the hydrocarbon industry [4]. Access to legacy data, particularly data that have been cleaned and standardised, remains a challenge, and exploratory data analysis is a vitally important first step ahead of supervised and unsupervised learning. The workflow presented herein could reduce uncertainty and aid decision making through the complete hydrocarbon value chain through exploration, appraisal, development and production. In addition, there is scope for legacy data and learnings from hydrocarbon basins to be repurposed to aid de-risking in emerging energy technologies, such as subsurface storage, carbon sequestration, and geothermal energy.

Author Contributions: R.E.B.: investigation, writing—original draft preparation and editing; visualisation. V.D.: supervision; conceptualization; methodology; validation; review. O.V.: data; conceptualization; review. R.N.: data; conceptualization; review. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Wood Mackenzie through funding of a Postdoctoral Research Associate position at Heriot Watt University, and through access to data from two basins.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are held by Wood Mackenzie.

Acknowledgments: This work was supported by Wood Mackenzie through funding research collaboration with Heriot-Watt University. All the data were anonymised and supplied by Wood Mackenzie and authors are thankful for the opportunity to publish the outcomes of this research. Authors also thank Mikhail Kanevski of University of Lausanne for the peer exchange on feature selection and the opportunities opened during his course on Machine Learning hands-on applications. Authors acknowledge the use of Orange Data Mining [27] and ML Office for SOM application [30]. We thank Susan Agar, who reviewed the paper most comprehensively and helped improve it along with two anonymous reviewers.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data, or in the decision to publish the results.

References

1. Vassiliou, M.S. *Historical Dictionary of the Petroleum Industry*; Rowman & Littlefield: Lanham, MD, USA, 2018.
2. Anand, P. Big Data is a big deal. *J. Pet. Technol.* **2013**, *65*, 18–21. [CrossRef]
3. Holdaway, K.R. *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data Driven Models*; Wiley: Hoboken, NJ, USA, 2014.
4. Perrons, R.K.; Jensen, J.W. Data as an asset: What the oil and gas sector can learn from other industries about "Big Data". *Energy Policy* **2015**, *81*, 117–121. [CrossRef]
5. Mayer-Schönberger, V.; Cukier, K. *Big Data: A Revolution that Will Transform How We Live, Work and Think*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.
6. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*; American Association for Artificial Intelligence: Palo Alto, CA, USA, 1996.
7. Standing, M.B. A pressure-volume-temperature correlation for mixtures of California oils and gases. *Drill. Prod. Pract. OnePetro* **1947**, *API 1947*, 275–287.
8. Vazquez, M.; Beggs, H.D. Correlations for fluid physical property prediction. In Proceedings of the SPE Annual Fall Technical Conference and Exhibition, Denver, CO, USA, 9–12 October 1977; OnePetro: Moscow, Russia, 1977; p. SPE-6719-M. [CrossRef]
9. Elsharkawy, A.M.; Alikhan, A.A. Correlations for predicting solution gas/oil ratio, oil formation volume factor, and undersaturated oil compressibility. *J. Pet. Sci. Eng.* **1997**, *17*, 291–302. [CrossRef]
10. Glasø, O. Generalized pressure-volume-temperature correlations. *J. Pet. Technol.* **1980**, *32*, 785–795. [CrossRef]
11. Al-Shammasi, A.A. A review of bubblepoint pressure and oil formation volume factor correlations. *SPE Reserv. Eval. Eng.* **2001**, *4*, 146–160. [CrossRef]
12. Tohidi-Hosseini, S.M.; Hajirezaie, S.; Hashemi-Doulatabadi, M.; Hemmati-Sarapardeh, A.; Mohammadi, A.H. Toward prediction of petroleum reservoir fluids properties: A rigorous model for estimation of solution gas-oil ratio. *J. Nat. Gas Sci. Eng.* **2016**, *29*, 506–516. [CrossRef]
13. Gharbi, R.B.; Elsharkawy, A.M. Neural network model for estimating the PVT properties of Middle East crude oils. *SPE Reserv. Eval. Eng.* **1999**, *2*, 255–265. [CrossRef]
14. Gharbi, R.B.; Elsharkawy, A.M.; Karkoub, M. Universal neural-network-based model for estimating the PVT properties of crude oil systems. *Energy Fuels* **1999**, *13*, 454–458. [CrossRef]

15. Ramirez, A.M.; Valle, G.A.; Romero, F.; Jaimes, M. Prediction of PVT properties in crude oil using machine learning techniques MLT. In Proceedings of the SPE Latin America and Caribbean Petroleum Engineering Conference, Buenos Aires, Argentina, 17–19 May 2017; OnePetro: Moscow, Russia, 2017. [CrossRef]
16. Oloso, M.A.; Khoukhi, A.; Abdulraheem, A.; Elshafei, M. Prediction of crude oil viscosity and gas/oil ratio curves using recent advances to neural networks. In Proceedings of the SPE/EAGE Reservoir Characterization & Simulation Conference, Abu Dhabi, UAE, 19–21 October 2009; European Association of Geoscientists & Engineers: Utrecht, The Netherlands, 2009; p. cp-170. [CrossRef]
17. Zadeh, L.A. Fuzzy sets. *Inf. Control.* **1965**, *8*, 338–353. [CrossRef]
18. Ali, J.K. Neural Networks: A New Tool for the Petroleum Industry? In Proceedings of the European Petroleum Computer Conference, Aberdeen, UK, 15–17 March 1994. [CrossRef]
19. Saemi, M.; Ahmadi, M. Integration of genetic algorithm and a coactive neuro-fuzzy inference system for permeability prediction from well logs data. *Transp. Porous Media* **2008**, *71*, 273–288. [CrossRef]
20. Karimpouli, S.; Fathianpour, N.; Roohi, J. A new approach to improve neural networks' algorithm in permeability prediction of petroleum reservoirs using supervised committee machine neural network (SCMNN). *J. Pet. Sci. Eng.* **2010**, *73*, 227–232. [CrossRef]
21. Tahmasebi, P.; Hezarkhani, A. A fast and independent architecture of artificial neural network for permeability prediction. *J. Pet. Sci. Eng.* **2012**, *86*, 118–126. [CrossRef]
22. Bhatt, A. Reservoir Properties from Well Logs using neural Networks. Ph.D. Thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2002.
23. Tewari, S.; Dwivedi, U.D.; Shiblee, M. Assessment of Big Data analytics based ensemble estimator module for the real-time prediction of reservoir recovery factor. In Proceedings of the SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 18–21 March 2019; OnePetro: Moscow, Russia, 2019.
24. Tahmasebi, P.; Hezarkhani, A. Application of adaptive neuro-fuzzy inference system for grade estimation; case study, Sarcheshmeh porphyry copper deposit, Kerman, Iran. *Aust. J. Basic Appl. Sci.* **2010**, *4*, 408–420.
25. Koren, Y.; Carmel, L. Visualization of labeled data using linear transformations. In Proceedings of the IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714), Seattle, WA, USA, 19–21 October 2003; pp. 121–128. [CrossRef]
26. Orange Data Mining. Linear Projection. 2015. Available online: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/visualize/linearprojection.html (accessed on 6 August 2021).
27. Demsar, J.; Curk, T.; Erjavec, A.; Gorup, C.; Hocevar, T.; Milutinovic, M.; Mozina, M.; Polajnar, M.; Toplak, M.; Staric, A.; et al. Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349–2353.
28. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
29. Longadge, R.; Dongre, S. Class imbalance problem in data mining review. *arXiv* **2013**, arXiv:1305.1707.
30. Kanevski, M.; Pozdnoukhov, A.; Timonin, V. *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*; EPFL Press: New York, NY, USA, 2009.
31. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of internal clustering validation measures. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, 13–17 December 2010; pp. 911–916.
32. Rokach, L.; Maimon, O. Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 321–352. ISBN 978-0-387-25465-4.
33. Kohonen, T. The self-organizing map. *Neurocomputing* **1998**, *21*, 1–6. [CrossRef]
34. Fix, E.; Hodges, J.L. *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*; Technical Report 4; USAF School of Aviation Medicine: Randolph Field, TX, USA, 1951.
35. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. Halotel, J.; Demyanov, V.; Gardiner, A. Value of geologically derived features in machine learning facies classification. *Math. Geosci.* **2020**, *52*, 5–29. [CrossRef]
37. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; IEEE: Picataway, NJ, USA, 1995; Volume 1, pp. 278–282.
38. Rodriguez-Galiano, V.F.; Sanchez-Castillo, M.; Dash, J.; Atkinson, P.M.; Ojeda-Zujar, J. Modelling interannual variation in the spring and autumn land surface phenology of the European forest. *Biogeosciences* **2016**, *13*, 3305–3317. [CrossRef]
39. Larter, S.; Wilhelms, A.; Head, I.; Koopmans, M.; Aplin, A.; Di Primio, R.; Zwach, C.; Erdmann, M.; Telnaes, N. The controls on the composition of biodegraded oils in the deep subsurface—Part 1: Biodegradation rates in petroleum reservoirs. *Org. Geochem.* **2003**, *34*, 601–613. [CrossRef]
40. Watson, K.M.; Nelson, E.F.; Murphy, G.B. Characterization of petroleum fractions. *Ind. Eng. Chem.* **1935**, *27*, 1460–1464. [CrossRef]
41. Bjørlykke, K. *Petroleum Geoscience*; Springer: Berlin/Heidelberg, Germany, 2015.
42. Feblowitz, J. Analytics in oil and gas: The big deal about big data. In Proceedings of the SPE Digital Energy Conference, The Woodlands, TX, USA, 5–7 March 2013; OnePetro: Moscow, Russia, 2013; p. SPE-163717-MS.