# Predicting Permeability Based On Core Analysis*

Harry Kontopoulos[1][0000−0002−9500−4906], Hatem Ahriz[1][0000−0002−1389−3886], Eyad Elyan[1][0000−0002−8342−9026], and Richard Arnold[2]

[1] School of Computer Science and Digital Media,The Robert Gordon University, Garthdee Rd, Aberdeen AB10 7QB
[2] Corex (UK) Ltd, Units B1 - B3, Airport Industrial Park, Howe Moss Drive, Dyce, Aberdeen AB21 0GL

**Abstract.** Knowledge of permeability, a measure of the ability of rocks to allow fluids to flow through them, is essential for building accurate models of oil and gas reservoirs. Permeability is best measured in the laboratory using special core analysis (SCAL), but this is expensive and time-consuming. This is the first major work on predicting permeability in the in the UK Continental Shelf (UKCS) based only on routine core analysis (RCA) data and a machine-learning approach. We present a comparative analysis of the various machine learning algorithms and validate the results, using permeability measured on 273 core samples from 104 wells. Results suggest that machine learning can predict permeability with relatively high accuracy. This opens new research directions in particular in the oil and gas sector.

**Keywords:** Machine Learning · Support Vector Regression · Core Analysis · Permeability prediction.

## 1 Introduction

A range of different data is generated during the life-cycle of oil and gas fields, from exploration to abandonment. This data include regional geology, seismic reports, sedimentological models, drilling data, fluid and rock properties [16]. Geologists, reservoir engineers and other scientists combine this data and use their expertise to construct models of the reservoir, evaluate the volume of available hydrocarbons and engineer the most efficient and profitable way to extract them from the reservoir [18]. The most direct type of geological data about the formation come from core analysis, the laboratory examination of well core samples extracted during the drilling. It is the only time scientists can see and physically examine material from within the reservoir.

Core analysis is usually divided into two stages. The first stage is called Routine Core Analysis (RCA) and the second stage is called Special Core Analysis

* This work is part of a Knowledge Transfer Partnership (KTP) programme, funded by Corex UK Ltd. and Innovate UK.

(SCAL) [16]. RCA usually includes tests such as fluid saturations, porosity and permeability measurements. Those measurements are taken on plugs or core samples. A SCAL programme might include the measurement of relative permeabilities, capillary pressures and wettability among others. Furthermore, the effect of coring and other fluids on the SCAL parameters could be used to evaluate the damage they cause to the formation[23].

Core analysis data are usually expensive and time consuming (several weeks) to obtain [27]. However, they are deemed to be the most accurate source of information for reservoir characterisation and a thoroughly designed core analysis programme can result in a more productive reservoir later in their lifetime [18].

## 2 Related work

Machine learning techniques have been used in the past in the context of core analysis mainly to extrapolate rarely available core analysis data to other more available types of data such as well log data [27]. Examples include prediction of permeability of gas reservoirs using well logs and core data [9] [22], identifying drilling sweet-spots for gas hydrate reservoirs without pre-existing well logs [15] [10], rock texture image classification using support vector machines [21], predicting permeability during acidizing [11] and predicting the optimal rate of penetration during drilling [12]. The work closest to ours is the study by Erofeev et al. [5] on the Chayandinskoye oil and gas condensate field in Russia. However, that study was limited to a single field and used desalination instead of drilling mud application during core analysis.

In contrast, this work relies on high quality data of actual permeability measurements obtained in the laboratory from a substantial number of core samples across a large number of wells across multiple fields of the UKCS and the north sea. Oil and Gas operators value the core analysis derived data as indispensable. However, budget constraints often limit the number of tests included in core analysis projects. A reliable and effective predictive method could be used alongside traditional routine core analysis techniques to fill the gap. The aim is to be able to provide the next best estimate when there is not enough funding for extensive laboratory measurements. As a proof of concept, the permeability is predicted after drilling mud application has been performed to the samples.

## 3 Methods

### 3.1 Dataset

The private dataset used here is part of the historical archive of Corex UK Ltd. It covers a significant part of the offshore area of the UK Continental Shelf (UKCS) and the north sea.

After data cleaning and preparation, the final dataset contained 273 observations and 13 features (Table 1). The number of samples might look small but core analysis is a laborious process with relatively small pace of generating data.

**Table 1.** Table showing the variables used, their corresponded type in R and their units

| Feature | Type | Units |
|---|---|---|
| Top Depth | Numeric | m |
| Pore Volume | Numeric | cc |
| Porosity | Numeric | rate |
| Grain Density | Numeric | gcc |
| Gas Permeability | Numeric | mD |
| Initial Permeability | Numeric | mD |
| Final Permeability | Numeric | mD |
| Brine Concentration | Numeric | ppm |
| Mud Weight | Numeric | ppg |
| Mud Type | Character | NA |
| Reservoir Temperature | Numeric | C |
| Pore Pressure | Numeric | atm |
| Overburden Pressure | Numeric | psi |

The features include the pore volume, porosity, grain density, the top depth of the core, gas permeability, initial permeability, final permeability (output), brine concentration, mud weight, mud type, reservoir temperature, pore pressure, overburden pressure. Mud type is a categorical feature and it is encoded into three dummy variables, with LTOBM (Low Toxicity OBM) the reference level.

Initial permeability is the permeability measurement before drilling mud application while final permeability is the permeability measured after drilling mud application. Drilling mud application for the context of this research means the laboratory simulation of the drilling procedure in the field using a specific drilling mud system. Drilling mud systems are expected to interact with the rock formation and potentially reduce its permeability (Figure 1). When centering or scaling of the input data is performed it is explicitly mentioned at the relevant model subsection otherwise the original values were used.

### 3.2 Prediction Models

A range of well established machine and statistical learning algorithms were applied to the given dataset. The main research question was whether the final permeability, after drilling mud application, can be predicted using the results of routine core analysis (RCA) tests as input. Therefore, the final permeability was selected as the output of the models. All the variables in table 1 were used as input parameters.

**Least Squares Linear Regression** The starting point of this research was to try and fit a linear model, represented by the formula:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \tag{1}$$

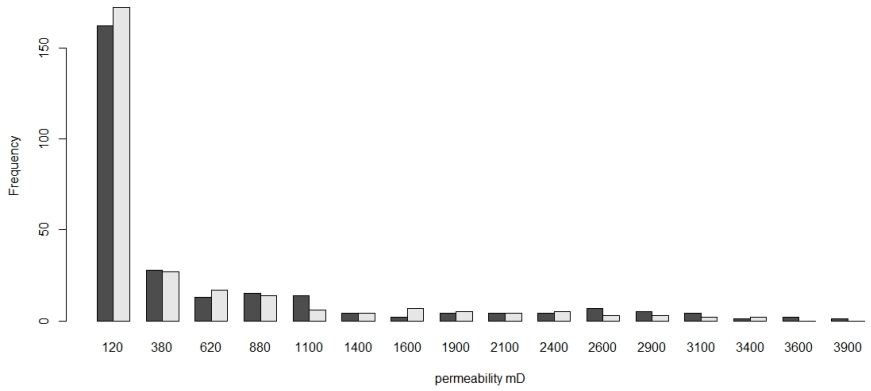


**Fig. 1.** The distribution of permeabilities before (dark grey) and after drilling mud application (light grey).

Least squares estimates the coefficients that minimise the following:

$$RSS = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta j x_{ij} \right)^2 \tag{2}$$

The linear model was fitted using the **lm()** function from the R stats package [19]. The QR decomposition of the input matrix is used to estimate the coefficients [6].

**Ridge Regression** Ridge regression adds a penalty to the loss function of least squares (equation 3.2). The penalty has the form $\lambda \sum_{j=1}^{p} \beta_j^2$ [13]. The penalty is controlled by the hyper-parameter $\lambda$, which is usually selected with grid search and cross validation. Ridge was fitted using the glmnet package in R. [7]

**Lasso Regression** This model is similar to Ridge except it applies an $\ell_1$ penalty to minimise the RSS subject to the constraint $\lambda \sum_{j=1}^{p} |\beta j|$ [24]. Lasso tries to address some of the problems arising in Ridge regression. The $\ell_2$ penalty used in Ridge minimises the coefficients towards zero but it does not turn any of them to exactly zero. Lasso instead can set a coefficient to zero and effectively perform feature selection. Lasso was fitted using the glmnet package in R[7].

**Partial Least Squares (PLS)** PLS identifies a set of components $Z_1, ...Z_M$ that are a linear combination of the original features [25]. The new components are fitted so they explain most of the variance in the predictors [14]. The idea behind this approach is that there are latent variables affecting the output that are not necessarily measured or captured in the dataset [26].

**Support Vector Regression (SVR)** Support Vector Regression estimates the weights of a hyperplane such that the RSS of the support vectors is minimised [4]. The support vectors are the only part of the dataset that participate in the estimation of the hyperplane equation and the minimization of the loss function. The input space was transformed into a higher dimension feature space using the radial basis function [3]. The SVR model was fitted using the e1071 R package [17]. The variables were scaled for zero mean and unit variance. The cost and gamma hyper-parameters were estimated by 10-fold cross validation, using the tune.svm() function.

**Artificial Neural Networks (ANNs)** A multi-layer perceptron [20] was used consisting of a feed-forward neural network with 13 neurons at the input layer, two hidden layers with five and three neurons respectively and an output layer with a single neuron for the final permeability. The model was trained using resilient back-propagation (RPROP) with weight backtracking [20] and the neuralnet R package [8]. The input was scaled with mean 0 and unit variance. The learning rate was set to 100 and the maximum number of allowed steps to 1e+05. The sum of square errors was the loss function.

**Decision Trees and Random Forests** Regression Trees [1] predict the value of a continuous variable by dividing the input space into $j$ distinct and not overlapping areas. For every new observation that falls into this area the average of the values of the training observations is returned. The regression tree was fitted with the tree package in R [2]. Random Forests build a number of trees in a bootstrapped version of the training samples. In each split step it only considers a random sample of $m$ predictors from the total available ones. This number is typically $\sqrt{p}$, where $p$ is the total number of features[14]. It then uses averaging across the trained trees to produce a final prediction.

## 4 Results and Discussion

A summary of the results for predicting the final permeability, after drilling mud application, using the various algorithms is presented in (Table 2). The dataset was divided into a training set (67%) and a test set (33%). The $R^2$, given by $1 - \frac{RSS}{TSS}$ with $TSS = \sum_i^n (y_i - \bar{y})$, on the training set and the MSE, given by $\frac{1}{n}\sum_i^n (\hat{y}_i - \bar{y})$, on the test set will be used to evaluate the models on the training and test set respectively.

$R^2$ scores show that many algorithms fit the training data well. SVR, Least Squares and PLS having the lowest MSE (Figure 2) on the test set. PLS and Lasso can give us great insight on the predictors affecting the response variable. According to the Least Squares model initial permeability, gas permeability and pore volume are the most significant features. (Figure 3).

Lasso produced a sparse model only assigning the pore volume, initial permeability and mud type(WBM) non zero coefficients. PLS also produced a sparse

**Table 2.** Table showing the $R^2$ on the training set and the MSE on the test set.

| Algorithm Name | $R^2$ | MSE |
|---|---|---|
| Least Squares | 0.916 | 50,020.24 |
| Linear Lasso | 0.725 | 50,927.96 |
| Linear Ridge | 0.881 | 51,251.47 |
| PLS | 0.894 | 50,150.91 |
| SVR | 0.900 | 44,135 |
| ANN | 0.997 | 323,477.3 |
| Decision Tree | 0.937 | 108,424.7 |
| Random Forest | 0.925 | 57,202.81 |

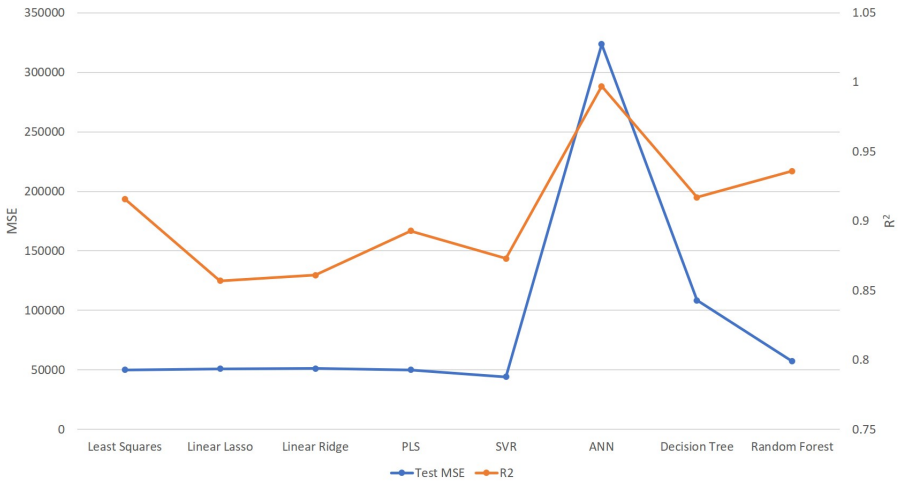


**Fig. 2.** Compare the models using $R^2$ on the training set (orange) and MSE on the test set(blue).

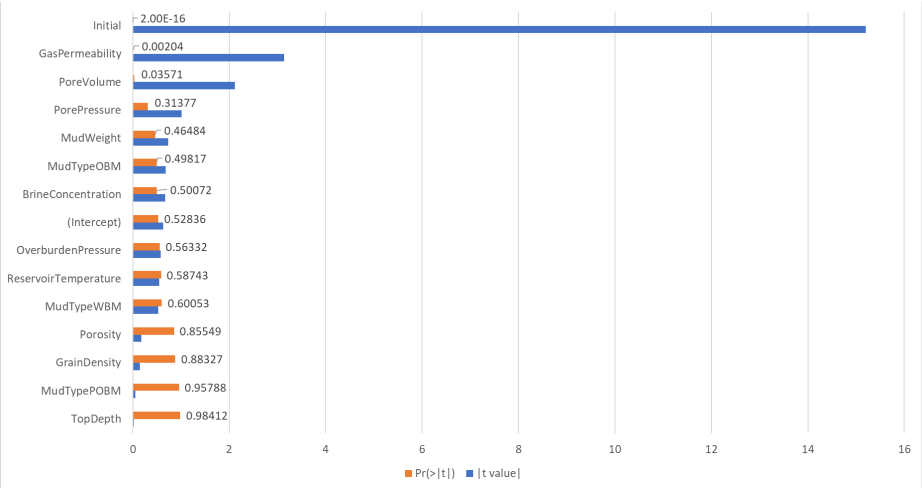


**Fig. 3.** The t value and the associated p value of the features for the least squares model.

model. The first six components are linear combinations of the top depth, gas permeability, initial permeability, brine concentration, pore pressure and overburden pressure.

**Linear Regression Least Squares** The linear regression model has a $R^2$ on the training set of 0.916 and test set MSE 50020.24 (Figure 4). $R^2$ suggests that the model fits the data relatively well but the MSE on the test set indicates that the predictive power requires further improvement.

**Linear Lasso** The best lambda for the Lasso model was estimated by 10-fold cross validation at 12.429. The non zero coefficients for the best $\lambda$ are pore volume, gas permeability, initial permeability, and mud type (WBM). Lasso has a test MSE of 50927.96 (Figure 5). The model does not fit the training data as well as the Least Squares but it still manages to generalise well on the test set according to the MSE figure.

**Linear Ridge** Ridge regression $\lambda$ parameter was estimated similarly to Lasso using 10-fold cross validation. The value that resulted in the lowest cross validation error was 89.749. Ridge regression performance on the training dataset was estimated by means of $R^2$ at 0.881 (Figure 5).

**Partial Least Squares** PLS fit the data on par with the regularised linear models and SVR. $R^2$ is estimated at 0.894 (Figure 6) on the training set and the

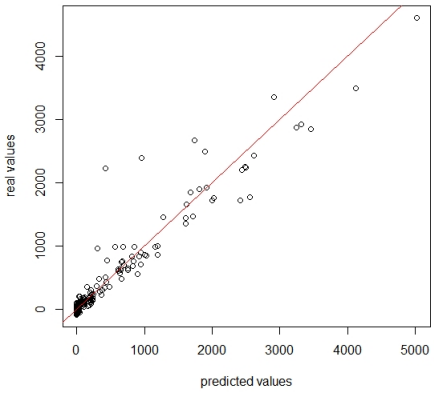


**Fig. 4.** The real against the predicted permeability values on the training set by Least Squares Regression.

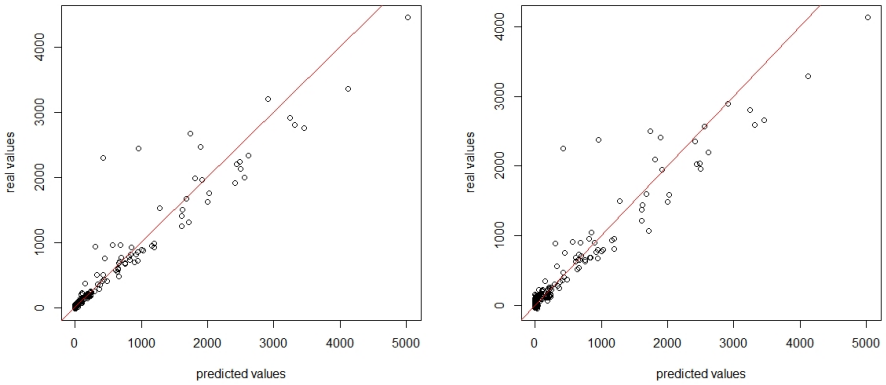


**Fig. 5.** The real against the predicted permeability values on the training set for the Lasso (left) and the Ridge model (right).

MSE at 50150.91 on the test set. The number of components of the final model was estimated by 10-fold cross validation. The number of components with the lowest CV error was 5 components and it was the one used to generate the model PLS $R^2$ and MSE values.

**Support Vector Regression** SVR gave the most promising results so far. The SVR hyper-parameters were estimated by grid search at $\gamma = 0.01$ and cost $= 10$ . The model fit the training data with a $R^2$ value of 0.9. Its MSE on the test set was estimated at 44135 (Figure 6). SVR performs much better than Least Squares in the test set.

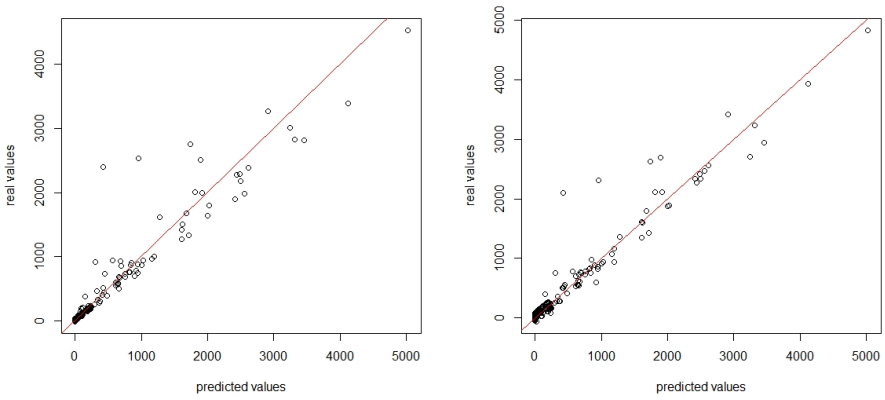


**Fig. 6.** The real against the predicted permeability values on the training set for the PLS (left) and the SVR model (right).

**Artificial Neural Network** The Artificial Neural Network was the model that followed the closest the training data with an $R^2$ value of 0.997. However, it performed very poorly on the test set with an MSE of 323477.3 (Figure 7). The high $R^2$ indicates that the model over-fits the training data while the very high test set MSE suggests that it fails to predict the permeability on unseen core samples.

**Decision Tree** The Decision tree model appeared to slightly over-fit the training data. Its $R^2$ was estimated at 0.937 and generalised poorly with MSE on test set at 57202.81 . The decision tree model performs better on the test set than the neural network but not as good as the SVR model.

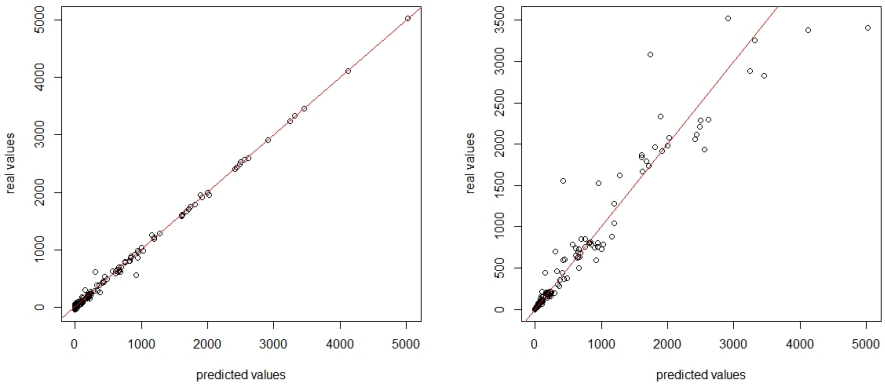


**Fig. 7.** The real against the predicted permeability values on the training set for the ANN (left) and the Random Forest model (right).

**Random Forest** The mtry hyperparameter, that controls the number of variables randomly sampled as candidates at each split, was estimated at 11 using 10-fold cross validation, repeated three times. The $R^2$ value on the training set was 0.925. The Random Forest model was an improvement compare to the decision tree model and generalised relatively well on the test set with an MSE of 57202.81 (Figure 7), albeit not as good as SVR.

## 5 Conclusion

This work showed that machine learning can be used alongside routine core analysis to predict final permeability, with SVR, Least Squares and PLS being the best models. Traditionally oil and gas companies only contract a small number of representative samples to be tested in the laboratory, due to financial constraints. Learning models based on historical data together with laboratory measurements of the limited number of financially approved measurements can alternatively provide a prediction for the rest of the available samples. This could be a new source of revenue for core analysis laboratories and a new service that can provide valuable information to operators to better manage their reservoirs. Future work might include more input features e.g. stratigraphic data that will improve the algorithm's performance in unseen cases or predict different types of output that can be of value for the oil and gas sector.

## References


1. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Taylor & Francis (Jan 1984)
2. Brian Ripley: tree: Classification and Regression Trees (2019), https://CRAN.R-project.org/package=tree, r package version 1.0-40
3. Burges, C.J.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery **2**(2), 121–167 (Jun 1998). https://doi.org/10.1023/A:1009715923555, https://doi.org/10.1023/A:1009715923555
4. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., Vapnik, V.: Support vector regression machines. In: Mozer, M.C., Jordan, M.I., Petsche, T. (eds.) Advances in neural information processing systems 9, pp. 155–161. MIT Press (1997), http://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf
5. Erofeev, A., Orlov, D., Ryzhov, A., Koroteev, D.: Prediction of Porosity and Permeability Alteration Based on Machine Learning Algorithms. Transport in Porous Media **128**(2), 677–700 (Jun 2019). https://doi.org/10.1007/s11242-019-01265-3, https://doi.org/10.1007/s11242-019-01265-3
6. Francis, J.G.F.: The QR Transformation A Unitary Analogue to the LR Transformation—Part 1. The Computer Journal **4**(3), 265–271 (Jan 1961). https://doi.org/10.1093/comjnl/4.3.265, https://academic.oup.com/comjnl/article/4/3/265/380632
7. Friedman, J.H., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software **33**(1), 1–22 (Feb 2010). https://doi.org/10.18637/jss.v033.i01, https://www.jstatsoft.org/index.php/jss/article/view/v033i01
8. Fritsch, S., Guenther, F., Wright, M.N.: neuralnet: Training of neural networks (2019), https://CRAN.R-project.org/package=neuralnet
9. Gholami, R., Shahraki, A.R., Jamali Paghaleh, M.: Prediction of Hydrocarbon Reservoirs Permeability Using Support Vector Machine (2012). https://doi.org/10.1155/2012/670723, https://www.hindawi.com/journals/mpe/2012/670723/
10. Gholami, R., Moradzadeh, A., Maleki, S., Amiri, S., Hanachi, J.: Applications of artificial intelligence methods in prediction of permeability in hydrocarbon reservoirs. Journal of Petroleum Science and Engineering **122**(C), 643–656 (2014). https://doi.org/10.1016/j.petrol.2014.09.007
11. GÜmrah, F., S. Sarkar, Y. A. Tasti, D. Erbas: Genetic Algorithm for Predicting Permeability During Production Enhancement by Acidizing. Energy Sources **23**(3), 245–256 (Apr 2001). https://doi.org/10.1080/00908310151133942, https://doi.org/10.1080/00908310151133942
12. Hegde, C., Gray, K.E.: Use of machine learning and data analytics to increase drilling efficiency for nearby wells. Journal of Natural Gas Science and Engineering **40**, 327–335 (2017). https://doi.org/10.1016/j.jngse.2017.02.019
13. Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics **42**(1), 80–86 (2000). https://doi.org/10.1080/00401706.2000.10485983, http://www.tandfonline.com/doi/abs/10.1080/00401706.2000.10485983
14. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics, Springer-Verlag, New York (2013). https://doi.org/10.1007/978-1-4614-7138-7, https://www.springer.com/gp/book/9781461471370

15. Lee, J., Byun, J., Kim, B., Yoo, D.G.: Delineation of gas hydrate reservoirs in the Ulleung Basin using unsupervised multi-attribute clustering without well log data. Journal of Natural Gas Science and Engineering **46**, 326–337 (Oct 2017). https://doi.org/10.1016/j.jngse.2017.08.007, http://www.sciencedirect.com/science/article/pii/S1875510017303104
16. McPhee, C., Reed, J., Zubizarreta, I.: Core analysis: a best practice guide. Developments in petroleum science ; volume 64, Elsevier, Amsterdam, Netherlands (2015)
17. Meyer, D., Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2019), https://CRAN.R-project.org/package=e1071, r package version 1.7-3
18. Ottesen, B., Hjelmeland, O.: 2008: The Value Added from Proper Core Analysis p. 12
19. R Core Team: R: A language and environment for statistical computing (ISBN 3-900051-07-0). R Foundation for Statistical Computing (2019), https://www.R-project.org/
20. Riedmiller, M.: Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms. Computer Standards & Interfaces **16**(3), 265–278 (Jul 1994). https://doi.org/10.1016/0920-5489(94)90017-5, https://linkinghub.elsevier.com/retrieve/pii/0920548994900175
21. Shang, C., Barnes, D.: Support vector machine-based classification of rock texture images aided by efficient feature selection. In: The 2012 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (Jun 2012). https://doi.org/10.1109/IJCNN.2012.6252634
22. Singh, S.: Permeability Prediction Using Artificial Neural Network (ANN): A Case Study of Uinta Basin. Society of Petroleum Engineers (Jan 2005). https://doi.org/10.2118/99286-STU, https://www-onepetro-org.ezproxy.rgu.ac.uk/conference-paper/SPE-99286-STU
23. Stiles, J.J., Hutfilz, J.: The use of routine and special core analysis in characterizing Brent Group reservoirs, U. K. North Sea. Journal of Petroleum Technology; (United States) **44**(6) (1992). https://doi.org/10.2118/18386-PA
24. Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B **58**, 267–288 (1994)
25. Wold, H.: 11 - Path Models with Latent Variables: The NIPALS Approach**NIPALS = Nonlinear Iterative PArtial Least Squares. In: Blalock, H.M., Aganbegian, A., Borodkin, F.M., Boudon, R., Capecchi, V. (eds.) Quantitative Sociology, pp. 307–357. International Perspectives on Mathematical and Statistical Modeling, Academic Press (Jan 1975). https://doi.org/10.1016/B978-0-12-103950-9.50017-4, http://www.sciencedirect.com/science/article/pii/B9780121039509500174
26. Wold, S.: Personal memories of the early PLS development. Chemometrics and Intelligent Laboratory Systems **58**(2), 83–84 (Oct 2001). https://doi.org/10.1016/S0169-7439(01)00152-6, http://www.sciencedirect.com/science/article/pii/S0169743901001526
27. Wong, K.W., Fung, C.C., Ong, Y.S., Gedeon, T.D.: Reservoir Characterization Using Support Vector Machines. In: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). vol. 2, pp. 354–359 (Nov 2005). https://doi.org/10.1109/CIMCA.2005.1631494