Journal of Petroleum Science and Engineering 212 (2022) 110333

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

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

Application of non-parametric statistical methods to predict pumpability of geopolymers for well cementing

Hassan Hamie [a], Anis Hoayek [b], Bassam El-Ghoul [c], Mahmoud Khalifeh [d,*]

[a] Vienna University of Technology, Institute of Energy Systems and Electrical Drives, Energy Economics Group (EEG), Austria
[b] Mines Saint-Etienne, Universite Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F- 42023, Saint-Etienne, France
[c] Phoenicia University, College of Engineering, Department of Petroleum Engineering, Lebanon
[d] Dept. of Energy and Petroleum Eng., Faculty of Science and Technology, University of Stavanger, Norway

ARTICLE INFO

Keywords:
Decision tree
Logistic regression
Plug and abandonment
Geopolymer consistency

ABSTRACT

As a potential alternative to Portland cement, geopolymers are getting wider acceptance in the scientific world. On a laboratory scale, the latter is being experimented repeatedly to extract valuable and valid results. To complement the experimental work and to make use of the data that resulted from previous experiments, statistical and mathematical models are developed. This article aims to anticipate test results, extract statistical relationships from measured properties, and therefore minimize the time and trials needed to run experiments in laboratories. Five independent input parameters are measured that cover the SiO_2/K_2O ratio, temperature, time, liquid to solid ratio and the total water content. For each set of these input variables, the consistency of geopolymers was obtained.

Two statistical models have been developed in this regard, the Decision Tree, which is a heuristic machine learning model, and the Logistic Regression which is a probabilistic model that calculates and estimates the probability for a certain mixture, at different time, temperature, and other independent variables, to reach a certain consistency threshold.

Both model results indicate sufficient performance, and the modelers can use such methods to predict the consistency (pumping time) trends of an untested geopolymer mixture. The results of our models are further validated by additional statistical tests, such as the receiver operating characteristic curve.

1. Introduction

Well cementing and Plug and Abandonment (P&A) operations consist of restoring natural barriers with artificial ones. Portland cement is the dominant material used for zonal isolation and well abandonment as barrier material. The established cement barrier should meet the requirements of local regulators or international independent organizations such as American Petroleum Institutes, and Norwegians Standards. There are a number of criteria for zonal isolation or well abandonment materials, including but not limited to, provide long-term integrity (eternal perspective), impermeable, non-shrinking, ability to withstand mechanical loads/impact, and ensure bonding to steel and formation (NORSOK D-010, 2013).

Despite its widespread usage, Portland cement still faces shortcomings like volume change, permeability, low ductility, and long-term durability concerns. These may result in leakage of fluid and numerous well integrity challenges. Researchers have been trying to develop alternative barrier materials to Portland cement to address the challenges associated with the performance of cement and minimizing concerns related to the emission of CO_2 during its manufacturing process. Among available alternative materials to Portland cement, one may refer to alkali activated based cement, cement modified with amorphous silicates, and geopolymers.

Geopolymers, also known as inorganic polymers, are mainly composed of aluminosilicate obtained from waste streams which react in presence of alkali silicate solution (Khalifeh et al., 2015). On a laboratory scale, geopolymers developed for zonal isolation have shown superior properties compared to Portland cement. Different researchers have analyzed and studied the mechanical, physical and chemical properties of geopolymers. Of these properties one may refer to viscosity, thickening time, bulk volume change, uniaxial compressive strength, sonic strength, shear bond and hydraulic bond strengths,

* Corresponding author. Department of Energy and Petroleum Engineering, Faculty of Science and Engineering, University of Stavanger, Norway.
E-mail address: Mahmoud.khalifeh@uis.no (M. Khalifeh).

https://doi.org/10.1016/j.petrol.2022.110333
Received 2 June 2021; Received in revised form 2 February 2022; Accepted 20 February 2022
Available online 22 February 2022

contamination with drilling fluids, and long-term durability of geopolymers at downhole conditions (Khalifeh et al., 2014, 2016; Sugumaran, 2015; Salehi et al., 2016, 2017, 2018; Liu et al., 2017, 2019; Olvera et al., 2019; Eid et al., 2021). The use of geopolymers is still at an early stage, with no field applications until today, and it is mainly conducted at laboratory scale. The latter needs to be conducted extensively, repeatedly, and with consistency to have valid and valuable results, and this consumes time and effort. Previous experimental results have shown that geopolymerization reaction is a function of different parameters (e.g., reactivity and particle size of precursors, type of hardener, molar ratio, liquid to solid ratio, total water content, pH value, temperature, total two capacity cations, etc.) and extensive sensitivity analysis has already been conducted by different researchers (Chamssine et al., 2021; Salehi et al., 2016; Khalifeh et al., 2016). Although controlling pumpability of geopolymers by use of admixtures have been achieved, this a time-consuming task to conduct sensitivity analysis by changing dosage of the admixtures. Therefore, predicting trend of consistency can help researchers to manage time and resources. In addition, this prediction can be used in field to find out the pumping time of these geopolymers.

The objective of this article is to predict workability of the geopolymers by using statistical and mathematical models to theoretically suggest the optimum mix design. The resultant optimized mixture can then be experimentally tested and verified. This article considers benchmarking against the experimental data on the rock-based geopolymers and excludes wider range of experimental data on other types of geopolymers.

2. Experimental procedure

2.1. Materials preparation

Rock-based geopolymeric precursors – Geopolymeric precursors consist of ground naturally occurring rock, slag and source of amorphous silica, mixed to normalize the chemical composition of the precursors. Reactivity, particle size distribution and chemical composition of the precursors have been extensively discussed by Alvi et al. (2020) and Khalifeh (2016). However due to the complexity of modeling, particle size and chemical composition of the precursors have not been included as input parameters.

Hardener – The geopolymeric binder known as hardener used in this study is potassium silicate solution. The used modular ratio ranges between 2.08 and 2.45.

Deionized water – Used to adjust the total water to total solid ratio.

2.2. Testing procedures

API recommended test procedure (API RP 10B-2, 2013) is followed to prepare the slurries, and for measuring thickening time, known also with other names such as pumpability and workability. Atmospheric consistometer (see Fig. 1) is used to measure the thickening time of the slurries. Consistency is the ability of slurry to be pumped or its resistance to flow, which is the resultant of the internal cohesion forces between the molecules found in the material. Table 1 presents mix design of the geopolymeric slurries used in this study, besides the five main parameters that have impact on rock-based geopolymers with fixed precursor content. The selected temperature ramp-up rate was 1 °C/min, and two bottomhole circulating temperatures (BHCT) were selected for this study: 50 and 65°**C**. All these experiments were repeated three times to ensure reliability and consistency of the measured data.

Fig. 1. API mixer and the shape of its blade (left & middle) and the atmospheric consistometer and its part (right).

From an operational point of view, slurry consistency is the main variable that engineers use to measure the pumpability of cementitious material and consequently its placeability. In the context of well cementing, a consistency of 40Bc or less is generally considered as pumpable and above it as risky to pump. In this document, the time required to reach 100 Bc will be referred to as the thickening time.

Figs. 2 and 3 show the measured consistency data at 50 and 65 °C of BHCT, respectively. By ramping up the temperature, the consistency decreases until a certain limit. This is regarded as the dissolution phase where reactive aluminosilicates are dissolved and transported. Then, consistency starts to increase from 20 to 40 Bc, this is the oligomerization phase. The consistency increase from 40 to 100Bc takes less than 2 min for the W-78 slurry at 50 °C. This is the polycondensation phase where all oligomers get linked and the slurry solidifies. When the molar ratio of the slurry is increased, the polycondensation rate decreases and therefore, the right-angle-set disappears (see test conducted at 65 °C). All the tests have been conducted three times to minimize errors and uncertainties originate from experimental part of this study.

3. Knowledge and background on modeling

Classical regression models postulate a relationship between independent (predictor) and dependent (response) variables, based on specific functions, distributions type, and parameter estimates. Such models rely on methods such as least square estimation and maximum likelihood to estimate the regression parameters. Least square method and maximum likelihood are two main methods used to estimate parameters from a random sample in the form of a multivariate regression analysis. Least square method is calculated by fitting a mathematical function to the points from a data set that has the minimal sum of the deviations squared (least square error). Maximum likelihood depends on the joint distribution function. Therefore, maximizing the likelihood function determines the parameters that are most likely to produce the observed data.

Table 1
Experimental data.

Data set	Modular ratio (SiO2/K2O)	Temperature[a] range (°C)	Time[b] (minutes)	Liquid to solid ratio	Total water content (g)
1	2.08 (W-78)	25 -> 50	22.7 -> 51.3	0.5542	265
2	2.16 (W-79)	25 -> 50	20.5 -> 99.4	0.5542	266
3	2.21 (W-80)	25 -> 50	20.2 -> 106.7	0.5543	266
4	2.45 (W-108)	20 -> 65	10.2 -> 100.3	0.585	232
5	2.45 (W-109)	20 -> 65	13.5 -> 100.6	0.585	228
6	2.45 (W-113)	20 -> 65	17.9 -> 99.9	0.577	223

[a] Continuous set of temperature profile that starts with 0 until it stabilizes and reaches a certain value over time. The increment that dictates the increase is over 1/1000 per increment.

[b] Time is stopped when the reaction reaches the thickening time (Time to reach 100 Bc).

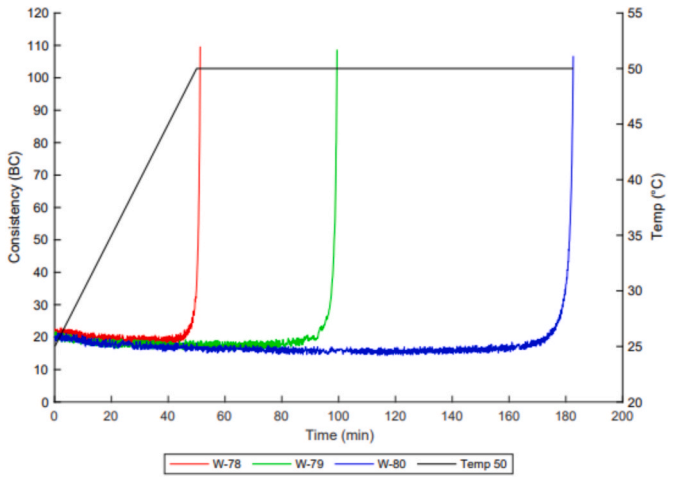

Fig. 2. Experimental data shown on graph with temperature profile (50 °C BHCT).

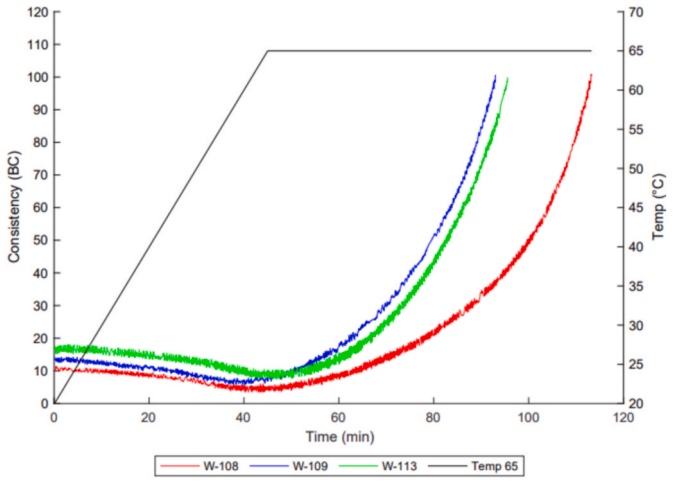

Fig. 3. Experimental data shown on graph with temperature profile (65 °C BHCT).

Other models look beyond linear regression (and variants) for complex multi-dimensional data sets. The idea is to extract the model from the data without making any assumptions regarding the underlying functional form, such models are referred to as machine learning. These methods are referred to as non-parametric, where no assumptions on data distribution are made. The latter models are also efficient in identifying patterns in data, capturing non-linear relationship between variables, and giving accurate results in classification problems, where response variable is categorical. Researchers started to include artificial intelligence tools to investigate the relationship between input parameters and choose the optimum compositions with minimal lab experiments (Van Dao et al., 2019; Pacheco-Torgal et al., 2014). The main advantages of machine learning are its ability to learn directly from the observed data, its ability to find patterns from incomplete data sets, and its ability to generalize results from out of sample observations (Nazari and Pacheco-Torgal, 2013).

Dao et al., (2019) used artificial intelligence to predict the geopolymer compressive strength based on the analysis of input parameters such as fly ash, sodium silicate solution, sodium hydroxide, and water content. They used two different machine learning models to predict the compressive strength. Such models are Artificial Neural Network (ANN) and Adaptive neuro-fuzzy inference system (ANFIS). The results showed that both techniques performed well and consequently show strong potential for predicting the compressive strength of geopolymer cement, and this is clearly indicated by the low mean absolute error and high coefficient of determination of both models, but with an edge that the ANFIS model has on ANN. On the other hand, other researchers were able to predict compressive strength of geopolymers by using ANN and changing its architecture and the choice of its input data. They were able to achieve decent results (Nazari and Pacheco-Torgal, 2013; Yadollahi et al., 2015; Siva Krishna and Ranga Rao, 2019). The input data used in each study are tabulated in Table 2.

Researchers have also considered using decision tree to predict the concrete strength, which resulted in a high accuracy prediction. The decision tree showed better results relative to the Linear Regression (LR), and ANN (Deepa et al., 2010). In addition to the compressive strength, researchers developed ANN modules to predict setting time and geopolymerization peak heat. Three models were conceived separately with SiO_2/Na_2O ratio in alkali silicate solution, alkaline solution concentration in the liquid phase, and the liquid/fly ash mass ratio (L/F), as variable input parameters (Ling et al., 2019). Although machine learning techniques have advantages regarding saving time, they still face several shortcomings. Where it requires a large amount of data, it also requires a lot of time for trial and error to reach the optimum algorithm.

Non-parametric models are vast and as previously mentioned; the most common method used is the Neural Network. However, it is not well suited in our context of study for four main reasons: First, the amount of data that is not large enough to let the neural network approach outperform other classical approaches. The more data one feeds a neural network, the better it gets. In other words, less data indicates weak approximation, which will eventually mean less capability of memorizing, and lead to less model performance when generalizing on a new data set. Second, most of the variables can be considered as categorical (see Table 2) and not a continuous set. As the possible values of some of the independent variables (liquid to solid ratio and total water content) belong to a finite set of values with low cardinality, so it is reasonable to consider these variables as categorical and to use them as an input for our decision tree. Third, using another traditional machine learning algorithm such as Decision Tree will be easier to interpret, especially when analyzing the results. Finally, we have noticed that most of cited researchers used machine learning ANN method, have constructed several network architectures and compared the result of the latter models. It is rather a drawback in such a method, as building a Neural Network is considered as a subjective study, where the choice of the number of layers, number of neurons, the nature of activation functions, stopping condition, among many other parameters, is not objective and several unjustified decisions have to be made in this regard. Therefore, and unless the user of the ANN model, conducts an exhaustive sensitivity analysis, and studies the robustness of the architecture chosen while constructing the model, it is meaningless to use neural networks and to consider a high level of complexity. The decision tree offers a better alternative, because of the nature of data which is a better fit for the structure of the Decision Tree algorithm, in addition there is no need for sensitivity, or trial and error estimates, which reduces the time of cross-validation.

The decision tree method will be complemented by a probabilistic analysis. Estimating the exact value of the consistency is not the primary objective of this analysis but the trend. In addition, our data is categorical and therefore the decision tree results will be complemented by evaluating the probabilistic distribution of the level of consistency. In other words, a second statistical approach is used to calculate and estimate the probability for a certain mixture, subjected to different physical and chemical properties, to reach a certain threshold of consistency. The latter method is called logistic regression, and it describes to us that when any of the input variables listed in Table 2 change, how the consistency will behave in a probabilistic way. This is very useful, especially, when there are limited experimental data.

Decision tree – It is a method based on data statistical machine

Table 2
Input statistical model variables found in the previous articles.

Authors	Nazari and Pacheco Torgal (2013)	Deepa et al. (2010)	Siva Krishna et al. (2019)	Van Dao et al. (2019)	Yadollahi et al. (2015)
Methods used	ANN	ANN, Linear regression, decision tree	ANN	ANFIS	ANN
Input variables	Curing time (days) Ca (OH) 2 content (wt%) Amount of superplasticizer (wt%) NaOH concentration (M) Mold type Geopolymer type H2O/Na2O molar ratio	Water to binder ratio Water content Fine aggregate ratio Fly ash replacement ratio Silica fume replacement ratio Super plasticizer	Molar concentration of alkali solution	Fly ash Sodium silicate solution Sodium hydroxide Water content	MS (SiO2/Na2O) Na2O content Water-blinder (w/b) ratio Ultrasonic pulse velocity
Output variable	Compressive strength of geopolymers	Compressive strength of concrete	Compressive strength of geopolymers	Compressive strength of geopolymers	Compressive strength of geopolymers

learning. Its operation is based on heuristics which, while satisfying the intuition, gives remarkable results in practice. It is an essential exploratory technique for uncovering structures in the data, and it is used to explain responses for a categorical set of variables. Its tree structure also makes them readable by a human being, unlike other approaches where the constructed predictor is a "black box".

Probabilistic analysis – It can calculate and estimate the probability for a certain mixture, at different times, temperatures, and other independent variables, to reach a certain threshold.

In this work, the decision tree method is applied to analyze lab results of the geopolymer pumpability, uncover structures in the data and explain response for a categorical set of variables, in order to predict the behavior of non-tested composition. Since the data is categorical, and to complement our analysis, a second statistical approach is used to calculate and estimate the probability for a certain mixture, subjected to different physical and chemical properties, to reach a certain threshold. The probability is calculated using a generalized multi-linear model with binary dependent variable, otherwise known as logistic regression.

4. Methods/data and models formulation

4.1. Model formulation of decision tree

A decision tree is formed by a set of branches and nodes. Starting from the root node to the leaf node each path represents a certain decision rule. The passage from a node to another is based on a logical "if-then" rule. In other words, a decision tree models a hierarchy of tests on the values of a set of variables called attributes. At the end of these tests, the predictor produces a numerical value or chooses an element from a discrete set of conclusions. In our context, we are in the case of regression because the value to predict is a real number (consistency of slurry).

To split a parent node into child nodes, we focus on the effect of each input variable on the target variable. So, starting by the root node one can split records into two or more categories based on characteristics that are related to the degree of homogeneity of the resultant child nodes. The most popular characteristic used to perform the splitting procedure is entropy, otherwise known as deviance or information attribute. It is analogous to the residual sum of squares in classical linear regression. Before explaining the algorithm, some notations will be presented in Table 3:

A node is called pure if all its observations belong to the same region of the dependent variable. Now, functions that make it possible to measure the degree of homogeneity in the different regions of the dependent variables are introduced. The most popular function used to measure the homogeneity is Entropy function defined by:

$$\mathscr{H}(p) = -\sum_{k=1}^{C} P(k|p)\ln P(k|p) \tag{1}$$

So, we recursively and as efficiently as possible divide the set of observations through tests defined using variables until one obtains subsets of observations containing (almost) only observations belonging to the same region (An entropy under a certain fixed threshold to define purity). The algorithm continues performing splitting until a pre-defined level of homogeneity is reached and this threshold is used as stopping criteria of the algorithm. This splitting procedure continues until predetermined homogeneity or stopping criteria are met. In our case, the target is to reach a status where all leaf nodes are pure with entropy of zero or where a pre-specified minimum change in purity cannot be made with any splitting methods. In some cases, a certain input variable is used more than one time in the splitting procedure, and in other cases some input variable is never used during path rule construction phase.

Our model was trained on five out of the six sets of data (see Table 2) and tested on the 6th set. The quota sampling approach is used instead of the purely random sampling approach to select the training and testing sets of data. That is why the first five sets out of six (representing 83% of the data) were selected to train the model and the last set to test it. This kind of approach is usually used when the underlying data is divided into several homogeneous subsets which is the case of our data. We selected the sets for training data in such a way as to get the best performance on the prediction level when moving to test data. The validation strategy is performed while taking into account the size of experimental data we have at hand. It is based on the fact that the data is divided into 5 sets for building both models and one set completely out of the bag to validate and test the predictive capability of the model. The selection of these sets is based on the principle of minimizing the out of bag errors (OOB errors) which is a technique widely used in machine learning to reduce overfitting.

An option to overcome the complexity of predicting exact values of a certain dependent output variable through a machine learning technique is to consider a probabilistic alternative approach based on the classical logistic regression, where the target is to compute the probability of being above a certain threshold instead of predicting the exact value.

4.2. Model formulation of logistic regression

To evaluate the probabilistic distribution of the level of consistency through a generalized multi-linear model with binary dependent variable, we define:

$$Y_i = F(\beta' X_i) + \varepsilon_i \; ; i \in \{1, 2, \ldots, N\} \tag{2}$$

where, N = number of observations, $F(\bullet)$ is a cumulative distribution function (CDF), ε_i is the residual term with $\mathbb{E}[\varepsilon_i] = 0$ and Y_i follows a Bernoulli distribution of parameter. The notations for the Logistic regression are found in Table 4:

Then,

$$Y_i = \begin{cases} 1 & \text{When the level of consistency is above a certain threshold} \\ 0 & \text{Otherwise} \end{cases}.$$

The vector $X_i = (X_{1i}, X_{2i}, \ldots, X_{pi})$ represents the p independent vari-

Table 3
Nomenclature of the decision tree.

Variable	Definition
n	Number of observations
k	Index of leaf node representing a particular region of the dependent variable (k = 1, …,153)
p	Position of the node inside the Tree
$N(p)$	Number of observations associated with the node that is in position p. The latter is related to the corresponding layer in the decision tree
$N(k\|p)$	Number of observations related to a certain region k of the dependent variable given that we are at the position p
$P(k\|p)$	Proportion of observations belonging to region k among those in position p
$\mathcal{H}(.)$	Shannon's Entropy

Table 4
Nomenclature of the logistic regression.

$$\pi_i = \mathbb{E}[Y_i] = \mathbb{P}[Y_i = 1|X_i] = F(\beta' X_i) \quad (3)$$

Variable	Definition
Y	Binary dependent variable
N	Number of observations
$F(\bullet)$	Cumulative Distribution Function, CDF
i	Index of observation
ε_i	Residual term
$\mathbb{E}[.]$	Expected value
π_i	Probability of observation i, to be above a certain threshold of consistency
X_i	Independent variable
β	Regression coefficients
β'	Transpose of B
$e(.)$	Exponential function
y_i	Observed values of the variables Y_i
x_i	Observed values of the variables X_i
$L(y_1,\ldots,y_N)$	Likelihood function
$l(y_1,\ldots,y_N)$	Log-likelihood function
$\widehat{\beta}$	Estimated parameters, by the maximum likelihood method
β_0	Intercept
h	Length of a step between two points
OR_i	Odds ratio of variable X_i

ables for the i^{th} observation, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of coefficients to be estimated. Therefore, the problem consists of estimating π_i based on observations X_i. In our study, the dependent variable is qualitative with binary outcomes, the most commonly used CDF is that of the logistic distribution defined as:

$$F(u) = \frac{e^u}{1+e^u} \quad (4)$$

That is the reason to call this regression, in this particular situation, "Binary Logistic Model" (Cox, 1958), where the probability, for observation$i$, to be above a certain level of consistency will be estimated by:

$$\pi_i = \frac{e^{\beta' X_i}}{1+e^{\beta' X_i}} \quad (5)$$

However, equation (5) is not useable unless one estimates the parametersβ. To do this, we use the classical principle of maximum likelihood. In our case, the likelihood function is given by:

$$L(y_1,\ldots,y_N) = \prod_{i=1}^{N} \mathbb{P}[Y_i = 1|x_i]^{y_i}(1 - \mathbb{P}[Y_i = 1|x_i])^{1-y_i}, \quad (6)$$

$$= \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

where, y_i and x_i are the observed values of the variables Y_i,X_i. By applying the logarithmic function, we will obtain the log-likelihood function:

$$l(y_1,\ldots,y_N) = \sum_{i=1}^{N}[y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)] \quad (7)$$

Thus, the estimation of β is done by maximizing the log-likelihood through solving the system of partial derivatives:

$$\frac{\partial l}{\partial \beta_j} = 0; j \in \{1,2,\ldots,p\} \quad (8)$$

Solution of system (8) is obtained by the iterative method of Newton-Raphson and the estimated parameters will be denoted by $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)$.

In our application, we consider the following independent variables: X_1 = "time in minutes", X_2 = "temperature in degree Celsius", X_3 = "Modular ratio", X_4 = "Liquid to Solid ratio (l/s)", and X_5 = "Total Water". On the other hand, the level of consistency is considered as dependent variable.

Therefore, an explicit expression of the logistic model is:

$$Y_i = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5) + \varepsilon_i; i \in \{1,\ldots,N\} \quad (9)$$

where, β_0 is the intercept coefficient.

Now, after estimating all the coefficients of the model and considering statistically significant ones, we can compute the probability that the consistency is above a certain priori fixed threshold for a given time, temperature, and ratio.

Analytically, these probabilities are given by:

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i}+\beta_4 X_{4i}+\beta_5 X_{5i}}}{1+e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i}+\beta_4 X_{4i}+\beta_5 X_{5i}}} \quad (10)$$

where, the index i is added to indicate the observation with$1 \le i \le N$.

Table 5
Non-parametric modeling results.

Machine Learning Model used	Decision Tree	ANN
Hyper-parameters	❖ The minimum sum of weights in a node in order to be considered for splitting (20) ❖ The minimum sum of weights in a terminal node (7) ❖ The depth of the tree (0) means that no restrictions are applied to tree sizes. ❖ The number of considered cross validations (10)	❖ Activation function = ReLU function ❖ Number of hidden layers = 1 ❖ Number of nodes in the hidden layer = 3 ❖ Loss function = Mean Squared Errors, MSE ❖ Optimizer = Adam ❖ Batch size = 50 ❖ Epochs = 200
R-squared	0.93	0.90
MSE	30.2	42.8

5. Results & discussion

5.1. Decision tree results

As mentioned in the above section, we have a set of 5 independent variables and the consistency as the dependent variable. In addition, there are six different compositions, otherwise defined as data sets (see Table 2). The aim is to analyze the data and try to establish a relationship between the variables, to anticipate the behavior of the geopolymers.

The results of the Decision Tree method are illustrated in Fig. 4. Starting from the root node, the temperature in our case, the second and third child nodes are formed by splitting it. The split point is the value of 64.917°C. As shown, a second independent variable is found in node 2 and it is the modular ratio (SiO2/K2O). The shape of the tree and the

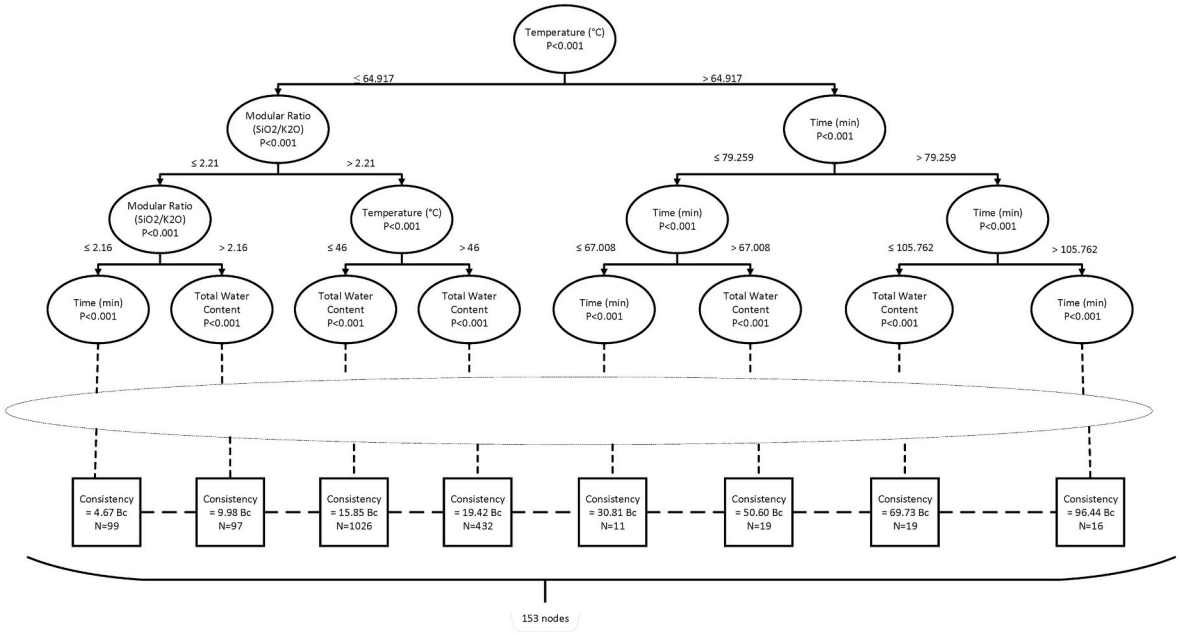

Fig. 4. Tree branches.

splitting procedure is about finding attribute that returns the highest information gain (lowest entropy). The final nodes, called leaf nodes, show the predicted outcome of consistency. It is worth noting that in our case we have 153 predicted outcomes for consistency and each of these discrete values is the mean value of observations in that node. For example, the final node on the left side of the graph has an average value of 4.67 *Bc* consistency; this value is the mean of 99 observations that fall under this node.

The decision tree in our case is graphically exhaustive, and it is important to note that certain input variables are used more than once in the partition (as can be seen in child node number 3, Temperature variables in that case).

Finally, we will have 153 leaf nodes consisting of different suggested values of consistency. These values will be used to make predictions of the consistency of a new set of observations by following the corresponding path throughout the tree.

Fig. 5 represents the real values of consistency for W-113 mixture and the predicted values. There are two predictive series. Once using the whole six data sets to construct the model and predict the consistency (In this case all the data set is used to train and test the data at the same time, otherwise known as in-sample forecasting), and the second is done using only a part of the data set (otherwise known as out-of-sample forecasting) to test and evaluate the model on the remaining data set. It is trivial to see that the model built on the whole sets of data is performing better than the model where we are splitting data between training and testing sets and this is mainly a consequence of an over fitting. To avoid an over fitting and to test the ability of the model to generalize and to give good predictions for new observations, it is better to adapt the out-of-sample approach.

To further visualize the results, Fig. 6 illustrates the predicted versus observed values. The lot resembles to a straight line at approximately 45°, and this is a clear indication of a good fit. This is further confirmed by calculating the correlation coefficient, which stands at a high value of 0.93. It is worth mentioning that we may have identical observed consistency values for different combinations of underlying independent variables, while the predicted value will differ, and this is particularly seen in Fig. 6, more specifically in the interval between 10 and 20.

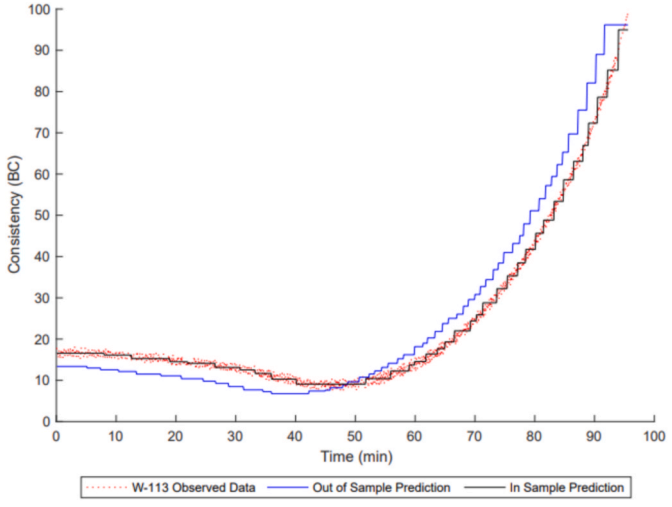

Fig. 5. Predictive power of the model.

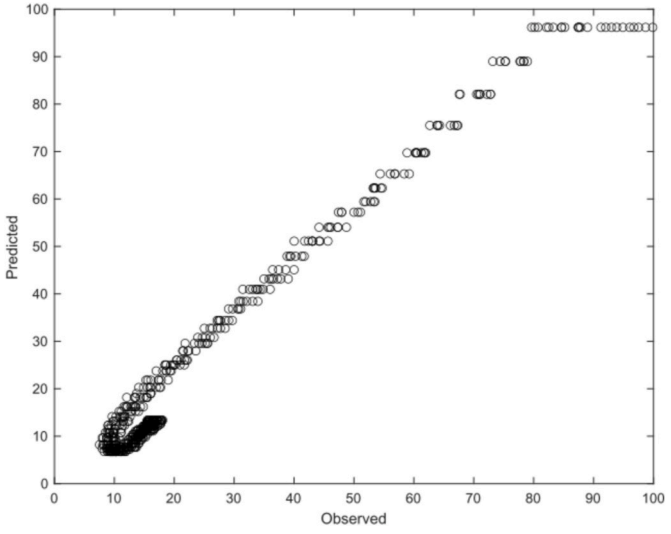

Fig. 6. Predicted vs. Observed.

Starting with one root node, the splitting of the data follows the path of more than 300 internal nodes "if then rule", until it reaches the final 153 leaf nodes. The splitting procedure that relies on entropy, used to measure the homogeneity of the data is efficient and reached the pre-specified target. In addition, the model can explain and evaluate the effects of the predictor variables (Temperature, time, liquid to solid ratio, total water content) one at a time, rather just all at once. This has been shown in Fig. 4. The tree structure makes the interpretation of the results easy. The comparison with the ANN model has been explained in theoretical terms, however and in order to apply both models on the same data set and compare the results, the authors decided to apply the ANN model. After applying the latter with optimal hyper-parameters, we got the following results:

To recap and to avoid the over-fitting issue in machine learning models, we are applying the following steps:

- Applying the classical technique of splitting the data into training and testing set. The model is built on the training set and tested on the testing set to assess its ability of generalization. The selection of the latter is based on an iterative cross validation.
- The complexity of the model was controlled. To do this control a regularization parameter was added to the objective (loss) function to penalize complex and deeper models.
- Finally, the main hyper-parameters of the neural network were tuned in a way to get an optimal result in terms of R2 and MSE on the testing data. To do this tuning a grid search technique (limited one, after considering some prior assumptions about some parameters, in order not to use all combinations) was applied.

5.2. Logistic regression results

Before applying the logistic model to the set of data, and to label our data into two binary values (two classes), we start by fixing a threshold for the consistency above which the binary dependent variable Y is equal to 1.

To do that, we will focus on approximating the increasing trend of the consistency around a certain consistency value fixed based on the previous graphical analysis. Based on a second order Taylor expansion and using numerical approximation, one can approximate the derivative of a function f at a point x_i by:

$$f'(x_i) \cong \frac{-f(x_{i+2}) + 4f(x_{i+1}) - 3f(x_i)}{2h} \tag{11}$$

where h is the length of the step between the points x_i, x_{i+1} and x_{i+2}. Then, by applying equation (11) we choose to approximate the increasing trend of consistency when the function reaches a level where the curve of the consistency in terms of temperature starts to increase in a significant way and the curve starts to a clear concave upward form. This number is selected to be the average of all the detected consistency levels for different ratios. The choice of the average is reasonable because the underlying data is well structured and there is no evidence of an outlier, which can disturb and skew the mean (see Table 5).

The results, for different values of ratios, are presented in Table 6. Based on the results presented in Table 6, we observe that the trend around a consistency level 15.7 is large enough to consider 15.7 as "*apriori*" threshold to build the binary dependent variable of the logistic regression model.

Now, by applying the maximum likelihood principle and considering the significant independent variables, we get the following estimators for the different parameters, see Table 7.

The negative sign of β_3 indicates that when the modular ratio increases the probability that the consistency is above 15.7 decreases, which fits with logically expected results. In addition, the positive sign of β_1 indicates that when the time increases the probability that the consistency is above 15.7 increases, which also fits with logically expected results.

Going beyond the results of the fitted logistic regression model and statistical significance, one can further investigate the impact of each predictor variable (see Table 8) with the dependent variable by exploring the odds ratio. The odds ratio of a variable X_i with a coefficient β_i is given by:

$$OR_i = e^{\beta_i} \tag{12}$$

The interpretation of OR_i can be as follows

- If $OR_i \cong 1$ this means that the variable X_i does not have a substantial impact on the decision that should be made on dependent variable level
- The more the OR_i exceeds unity, the more we can say that when the value of the variable X_i increases the probability of getting $Y = 1$ increases as well
- By the same token, we can say that if $0 \leq OR_i < 1$, the more we can say that when the value of the variable X_i decreases the probability of getting $Y = 1$ increases, inverse proportions.

It is clear that variables with the highest impact on the consistency are "Liquid to solid ratio" and "Modular Ratio". On the contrary, variables "Time" and "Temperature" do not have a big impact.

Now, one can compute the probability of a consistency level greater than 15.7 for any value given time an assumption about the independent variables by applying the following equation:

$$\pi = \frac{e^{427.16+0.35X_1+\ldots-1.58X_5}}{1+e^{427.16+0.35X_1+\ldots-1.58X_5}} \tag{13}$$

A good logistic regression model is the one that maximize the percentage of positives (i.e., the cases where $Y_i = 1$) that are successfully classified as positive (called specificity) and the percentage of negatives (i.e., the cases where$Y_i = 0$) that are successfully classified as negatives (called sensitivity). Table 9, shows the validation results and can be interpreted as such:

- Letter A, designates the successful prediction of the number of observations that are below the Consistency threshold.
- Letter B, designates the unsuccessful prediction of the number of observations that are above the Consistency threshold, and this number accounts for 433 observations. If our model was completely successful, then this number should be zero. In other words, there should be zero number of unsuccessful predictions.

Table 6
Taylor expansion results.

Ratio	2.08	2.16	2.21	2.45 (W-108)	2.45 (W-109)	2.45 (W-113)
Consistency level	21.01	18.33	16.17	7.37	17.67	13.63
Derivative $f'(x)$ approximation	37.87	30.178	34.79	45.21	39.26	48.40

Table 7
Maximum likelihood results.

Coefficients	Estimate	Std. Error	z value	Pr (>|z|)
Intercept (β_0)	110.27	5.993	18.398	<2e-16 ***
X_1 (*Time*, β_1)	0.015	0.001	11.535	<2e-16 ***
X_2 (*Temperature* β_2)	0.014	0.004	3.804	<1e-4 ***
X_3 (*Modular ratio* β_3)	−59.11	3.438	−17.191	<2e-16 ***
X_4 (*l/s ratio* β_4)	142.81	13.594	10.505	<2e-16 ***
X_5 (*Total water* β_5)	−0.222	0.014	−16.287	<2e-16 ***

***Indicates statistical significance with a confidence level of 0.1% level.

Table 8
Odd Ratio results.

Coefficients	Odd ratio
Time, β_1	$e^{\beta_1} = 1.015$
Temperature β_2	$e^{\beta_2} = 1.014$
Modular ratio β_3	$e^{\beta_3} \cong 0$
l/s ratio β_4	$e^{\beta_4} \cong \infty$
Total water β_5	$e^{\beta_5} = 0.8$

Table 9
Model Validation results.

Observed values of *Y*	Predicted values of *Y*		
	0	1	Total
0	2171 (A)	433 (B)	2604
1	1771 (C)	3246 (D)	5017
Total	3942	3697	7621

- Letter C, means that the model is unsuccessful in predicting the number of observations that are below the threshold, and therefore has detected 1771 observations.
- Letter D, designates the successful prediction of our model, for 3246 observations that are above the threshold.

Then, for this particular threshold, the sensitivity attributed to letter A, is $^{2,171}/_{2,604} \approx 83\%$ and the specificity is $^{3,246}/_{5,017} \approx 65\%$. However, a more powerful criterion to evaluate the global performance of a logistic regression without fixing any threshold is the Receiver Operating Characteristic (ROC) curve. ROC Curve gives us an idea of the performance of the model under all possible values of threshold by plotting Sensitivity in terms of specificity. The model is as good as the curve is close to the optimal situation (i.e., the point of coordinates$(1,1)$). In our context, and as can be seen in Fig. 7, the area under the ROC curve is close to 1, more specifically 0.88. Then, we can say that our model has a global prediction accuracy of88%.

The results of the fitted logistic regression model indicate that all the variables used to predict the consistency results are statistically significant. This confirms the dependency between all our data set. Going further and to measure the impact of each predictor variable with the dependent variable the odds ratios are computed. Results clearly show that variables with the highest impact on the consistency are "Liquid to solid ratio" and "Modular Ratio". Thus, when the value of both ratios increases, the probability of having a consistency level of 15.7 and above increases (threshold). To validate the probabilistic model, the sensitivity and specificity are computed. The results confirm that the model probability results can account successfully for most of the dataset and that few observations are overlooked. The latter, however, is a validation of the results for a particular threshold, and a more powerful criterion to test the global performance of the model is the ROC curve. While using the latter validation method we can have a good performance of the model under all possible values of threshold.

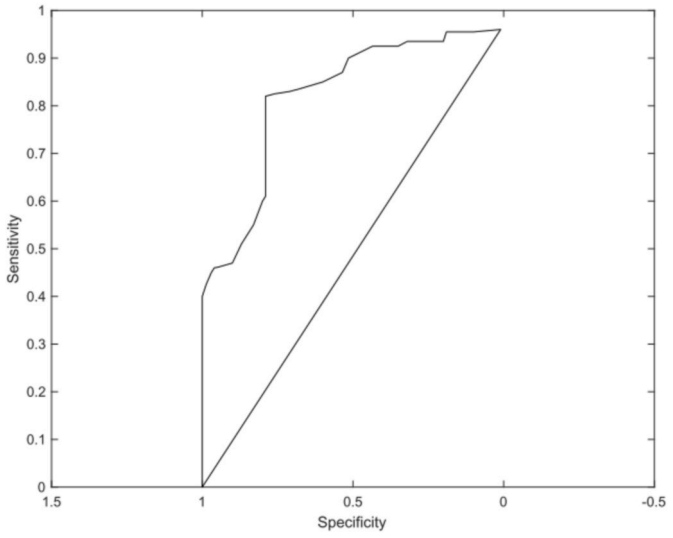

Fig. 7. ROC curve of Logistic Model.

6. Conclusion

Applying statistical models to experimental data can be useful, especially when the main aim is to understand the causal relationship of such variables. In this work, mathematical and statistical methods were applied to forecast consistency profile of the geopolymeric slurries, designed for downhole applications. Logistic regression is used as a classification method (compute probability), whereas the decision tree is used for forecasting purposes. The non-parametric Decision Tree, as an alternative to the commonly used Neural Network, was successfully used. The decision tree model could accurately forecast the consistency values. Accuracy of the forecast was illustrated by the high correlation of determination and an almost linear q-q plot (Fig. 6). The model could explain and evaluate the effects of the predictor variables (temperature, time, molar ration, liquid to solid ratio, total water content) one at a time, rather just all at once.

Fitted logistic regression model indicates that all the variables used to predict the consistency results are statistically significant with a considerable confidence level. Results show that variables with the highest impact on the consistency could be "liquid to solid ratio" and "modular ratio". Model results were further validated by the high global prediction accuracy.

To the best of the author's knowledge, it is the first time that such two models are combined in a complimentary way, in an attempt to study the impact and importance of independent variables on geopolymer consistency. One is used for point regression prediction, and the second is used for classification, and both confirm the same results. The second major contribution of this manuscript is that the authors are proposing a mathematical/numerical method based on Taylor expansion to fix the optimal threshold of the logistic regression while classical methods were fixing this threshold either based on simple descriptive statistical analysis or using expert's opinion. Finally, this study enables its user to perform sensitivity analysis by changing parameters, analyzing the modeling results, and then do experiments to validate the prediction. Time and resources management are the core drivers for conducting the project.

Although the limited data sets were sufficient for forecasting results, it is well known that a larger data set plays in a critical role in such statistical models, as long as the data contain meaningful information and not only noise. Therefore, and pursuing further research, one can enlarge the data set, in terms of the number of observations and the number of measured properties. In addition, other machine learning models can be used to model and predict the consistency, such as but not limited to, Random Forest, Bagging and Boosting techniques. Additionally, for future work, when more variables are available one can study the sensitivity of the results in different ways e.g., test the impact of each variable by a permutation of its values or bootstrapping techniques.

These models can also generalize the decision tree approach, and therefore we can extract more information in the context of supervised models with high complexity of the underlying independent variables.

Credit author statement

Mahmoud Khalifeh: Conceptualization, Methodology, Experimental work, Funding, Writing- Original draft preparation, Writing- Reviewing

and Editing, revising. Hassan Hamie: Conceptualization, Methodology, Data curation, Modelling, Writing- Original draft preparation, Writing-Reviewing and Editing, Revising. Bassam El-Ghoul: Modelling, Writing-Reviewing and Editing, Data curation, revising. Anis Hoayek: Conceptualization, Methodology, Data curation, Modelling, Writing- Original draft preparation, Writing- Reviewing and Editing.

Declaration of competing interest

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

Acknowledgement

The authors gratefully acknowledge TotalEnergies, AkerBP, ConocoPhillips and Research Council of Norway for financially supporting the SafeRock KPN Project (RCN #319014) at the University of Stavanger, Norway. Special thanks to Laurent Delabroy, Johan Kverneland, Roy Gordon Middleton for their technical inputs.

References

Alvi, M.A.A., Khalifeh, M., Agonafir, M.B., 2020. Effect of nanoparticles on properties of geopolymers designed for well cementing applications. J. Petrol. Sci. Eng. 191, 107–128. https://doi.org/10.1016/j.petrol.2020.107128, 2020.

Api Rp 10B-2, 2013. API RP 10B-2. Recommended Practice for Testing Well Cements, second ed.

Chamssine, F., Khalifeh, M., Eid, E., Minde, M.M., Saasen, A., 2021. Effects of temperature and chemical admixtures on the properties of rock-based geopolymers designed for zonal isolation and well Abandonment. In: Paper OMAe2021-60808 Published at the 40th International Conference on Ocean, Offshore and Arctic Engineering OMAE2021 June 21-30, 2021. Online).

Cox, D.R., 1958. The regression analysis of binary sequences. J. Roy. Stat. Soc. B 20 (2), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.

Deepa, C., SathiyaKumari, K., Sudha, V.P., 2010. Prediction of the compressive strength of high performance concrete mix using tree based modeling. Int. J. Comput. Appl. 6 (5), 18–24. https://doi.org/10.5120/1076-1406.

Eid, E., Tranggono, H., Khalifeh, M., Salehi, S., Saasen, A., 2021. Impact of drilling fluid contamination on performance of rock-based geopolymers. SPE J. https://doi.org/10.2118/205477-PA. Paper Number: SPE-205477-PA.

Khalifeh, M., 2016. Materials for Optimized P&A Performance: Potential Utilization of Geopolymers. Published at the University of Stavanger, Norway. PhD thesis UiS;292. https://uis.brage.unit.no/uis-xmlui/handle/11250/2396282.

Khalifeh, M., Hodne, H., Saasen, A., Obara, I., Eduok, E., 2016. Usability of geopolymers for oil well cementing applications: reaction mechanisms, pumpability, and properties. In: Paper SPE-182354-MS Presented at the SPE Asia Pacific Oil & Gas Conference and Exhibition, 25-27 October, Perth, Australia. https://doi.org/10.2118/182354-MS.

Khalifeh, M., Saasen, A., Vrålstad, T., Hodne, H., 2014. Potential utilization of geopolymers in plug and abandonment operations. In: Paper SPE-169231-MS Presented at the SPE Bergen One Day Seminar, 2 April, Bergen, Norway. https://doi.org/10.2118/169231-MS.

Khalifeh, M., Saasen, A., Vrålstad, T., Larsen, H.B., Hodne, H., 2015. Cap rock restoration in plug and abandonment operations; possible utilization of aplite-based geopolymers for permanent zonal isolation and well plugging. In: Paper SPE-175457-MS Presented at the SPE Offshore Europe Conference and Exhibition. https://doi.org/10.2118/175457-MS.

Ling, Y., Wang, K., Wang, X., Li, W., 2019. Prediction of engineering properties of fly ash-based geopolymer using artificial neural networks. Neural Comput. Appl. 13 https://doi.org/10.1007/s00521-019-04662-3, 2019.

Liu, X., Nair, S., Aughenbaugh, K., van Oort, E., 2019. Mud-to-cement conversion of non-aqueous drilling fluids using alkali-activated fly ash. J. Petrol. Sci. Eng. 182, 106–242. https://doi.org/10.1016/j.petrol.2019.106242, 2019.

Liu, X., Ramos, M.J., Lee, H., Espinoza, D.N., van Oort, E., 2017. True self-healing geopolymer cements for improved zonal isolation and well abandonment. In: Paper SPE-184675-MS Presented at the SPE/IADC Drilling Conference and Exhibition, 14-16 March, the Hague, The Netherlands. https://doi.org/10.2118/184675-MS.

Nazari, A., Pacheco Torgal, F., 2013. Predicting compressive strength of different geopolymers by artificial neural networks. Ceram. Int. 39 (3), 2247–2257. https://doi.org/10.1016/j.ceramint.2012.08.070.

Norsok D-010, 2013. Well integrity in drilling and well operations. Rev 4 (Standard Norway).

Olvera, R., Panchmatia, P., Juenger, M., Aldin, M., van Oort, E., 2019. Long-term oil well zonal isolation control using geopolymers: an analysis of shrinkage behavior. In: Paper SPE-194092-MS Presented at the SPE/IADC International Drilling Conference and Exhibition, 5-7 March, the Hague, The Netherlands. https://doi.org/10.2118/194092-MS.

Pacheco-Torgal, F., Labrincha, J.A., Leonelli, C., Palomo, A., Chindaprasirt, P., 2014. Handbook of Alkali-Activated Cements, Mortars and Concretes. Published by Elsevier, ISBN 978-1-78242-288-4.

Salehi, S., Ezeakacha, C.P., Khattak, M.J., 2017. Geopolymer cements: how can you plug and abandon a well with new class of cheap efficient sealing materials. In: Paper SPE-185106-MS Presented at the SPE Oklahoma City Oil and Gas Symposium, 27–31 March, Oklahoma City, Oklahoma, USA. https://doi.org/10.2118/185106-MS.

Salehi, S., Khattak, M.J., Ali, N., Rizvi, H.R., 2016. Development of geopolymer-based cement slurries with enhanced thickening time, compressive and shear bond strength and durability. In: Paper SPE-178793-MS Presented at the IADC/SPE Drilling Conference and Exhibition, 1-3 March, Fort Worth, Texas, USA. https://doi.org/10.2118/178793-MS.

Salehi, S., Khattak, M.J., Ali, N., Ezeakacha, C., Saleh, F.K., 2018. Study and use of geopolymer mixtures for oil and gas well cementing applications. J. Energy Resour. Technol. 140 (1), 012908 https://doi.org/10.1115/1.4037713.

Siva Krishna, A., Ranga Rao, V., 2019. Strength prediction of geopolymer concrete using ANN. Int. J. Recent Technol. Eng. 7 (6C2), 661–667.

Sugumaran, M., 2015. Study on effect of low calcium fly ash on geopolymer cement for oil well cementing. In: Paper SPE-176454-MS Presented at the SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, 20-22 October, Nusa Dua, Bali, Indonesia. https://doi.org/10.2118/176454-MS.

Van Dao, D., Ly, H.B., Trinh, S.H., Le, T.T., Pham, B.T., 2019. Artificial intelligence approaches for prediction of compressive strength of geopolymer concrete. Materials 12 (6), 983. https://doi.org/10.3390/ma12060983, 2019.

Yadollahi, M.M., Benli, A., Demirboğa, R., 2015. Prediction of compressive strength of geopolymer composites using an artificial neural network. Mater. Res. Innovat. 19 (6), 453–458. https://doi.org/10.1179/1433075X15Y.0000000020.