Water cut/salt content forecasting in oil wells using a novel data-driven approach

Rouhollah Ahmadi¹, Jamal Shahrabi^{2,*}, and Babak Aminshahidy¹

¹ Department of Petroleum Engineering, Amirkabir University of Technology, Tehran Polytechnique, No. 350, Hafez Ave, Valiasr Square, 1591634311 Tehran, Iran

² Department of Industrial Engineering & Management Systems, Amirkabir University of Technology, Tehran Polytechnique, No. 350, Hafez Ave, Valiasr Square, 1591634311 Tehran, Iran

Received: 6 September 2018 / Accepted: 20 June 2019

Abstract. Water cut is an important parameter in reservoir management and surveillance. Unlike traditional approaches, including numerical simulation and analytical techniques, which were developed for predicting water production in oil wells based on some assumptions and limitations, a new data-driven approach is proposed for forecasting water cut in two different types of oil wells in this article. First, a classification approach is presented for water cut prediction in sweet oil wells with discontinuous salt production patterns. Different classification algorithms including Support Vector Machine (SVM), Classification Tree (CT), Random Forest (RF), Multi-Layer Perceptron (MLP), Linear Discriminant Analysis (LDA) and Naïve Bayes (NB) are investigated in this regard. According to the results of a case study on a real Iranian sweet oil well, RF, CT, MLP and SVM can provide the best performance measures, respectively. Next, a Vector Autoregressive (VAR) model is proposed for forecasting water cut in salty oil wells with continuous water production during the life of the well. The proposed VAR model is verified using data of two real salty oil wells. The results confirm that the well-tuned proposed VAR model could provide reliable and acceptable results with very good accuracy in forecasting water production for the near future days.

Acronyms

ACF	Autocorrelation Function
AR	Autoregressive
BSW	Basic Sediment & Water
CT	Classification Tree
CUMPROD	Cumulative oil production between two successive days of shutting in the well due to rises in salt production
CV	Cross Validation
DDM	Data-Driven Modeling
DW	Distance from bottom of production interval of the well to the water-oil-contact
LDA	Linear Discriminant Analysis
MAE	Mean Absolute Error
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
OPR	Oil Production Rate

PACF	Partial Autocorrelation Function
RF	Random Forest
RMSE	Root Mean Squared Error
SAOW	Salty Oil Well
SCT	Salt Content
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
SWOW	Sweet Oil Well
VAR	Vector Autoregressive
WC	Water Cut

1 Introduction

The extraction of oil and gas from underground reservoirs is often accompanied by water or brine, which is referred to as produced water. As the reservoirs mature, especially if secondary or tertiary recovery methods are used, the quantity of produced water climbs and often exceeds the volume of the hydrocarbons before the reservoir is depleted. Produced

* Corresponding author: jamalshahrabi@aut.ac.ir;
jamalshahrabi@gmail.com

water is by far the largest volume byproduct stream of oil and gas exploration and production. The cost of producing, handling, and disposal of the produced water often defines the economic lifetime of a field and the actual hydrocarbon reserves; therefore, understanding and predicting the aspects, behavior, and problems induced by the produced-water flow is important (Petrowiki, 2018).

The major constituents of concern in produced water are Salt Content (SCT), oil and grease, inorganic and organic compounds, and naturally occurring radioactive material (Clark and Veil, 2009). Due to the increasing volume of waste all over the world in the current decade, the outcome and effect of discharging produced water on the environment has lately become a significant issue of environmental concern (Nasiri et al., 2017).

The cost of managing produced water is a significant factor in the profitability of oil and gas production. In addition, the cost-effective techniques for managing oilfield water require a comprehensive understanding of reservoir characteristics, production volumes, hydrogeology, engineering design, and environmental considerations (SLB, 2018). Therefore evaluating and predicting the volume of produced water in oil/ gas production wells could be considered as a major first step in planning for the future requirements of salty oil treatment in desalting facilities and the management of underground water resources as well.

Generally, the oil producing wells could be categorized into two different groups based on the entrained volume and patterns of producing water along with the oil, as listed in the following statements:

1. *Sweet Oil Well (SWOW)*: This type of well produces (sweet) oil with no excess salt or basic sedimentary; so the well flows to a production unit without needing any desalting processes. Some SWOWs are however encountered irregularly by sudden increases in salt production during their production life due to water coning phenomenon. To remediate this situation, the well may be shut-in or its production rate may be decreased for some days so that the cone of water rose in the formation around the wellbore recedes back to its original level in the aquifer. Figure 1 represents the abnormal behavior of salt production with rare and sudden peaks (defined as anomalies) over time in a daily scale for a real SWOW in Southwest Iran.

2. *Salty Oil Well (SAOW)*: This kind of well produces brine continuously with different volumes along with the oil. Therefore, the produced oil should always be treated in desalting facilities after processing in the production unit. Figure 2 illustrates the daily BSW production trends of two real oil wells, located in Southwest Iran, producing with varying WCs. Regarding the plots, the wells show a continuous production of brine with varying production rates along with the oil. No specific attentions have been made to WC prediction in SWOWs in the literature and no model is developed by other researches for predicting the WC or anomalous conditions in this type of well. This is, however, quite different for SAOWs where there are currently two main approaches for predicting Water Cut (WC) in them. The first is numerical simulation technique and another is analytical or empirical methods (Ershaghi and Omerigie,

Fig. 1. Production history of BSW content (in %) of a real SWOW in Southwest Iran collected over time (in a daily scale).

1978; Lawal et al., 2007; Li et al., 2011; Liu et al., 2017; Lo et al., 1990; Sitorus et al., 2006). The numerical simulation is time consuming and expensive to reach and requires information that is not readily accessible. Most of the analytical models, however, represent WC as a function of cumulative oil production instead of production times and are based on simplifying assumptions causing unacceptable results in complex reservoirs.

Only monotonic trends of production could be handled effectively by the empirical models as they are based on some predefined equations like exponential functions that are not capable of presenting multiple increasing or decreasing trends. Some wells may encounter cyclic, unstable or changing production patterns over time. These situations could not be captured very well using the traditional models as well. These methods also utilize a fixed set of input variables in their structure and are not flexible in using more or less variables, if needed.

Recent developments in computational intelligence, in the area of machine learning in particular, have greatly expanded the capabilities of empirical modeling in all areas of science and industry. The field which encompasses these new approaches is called Data-Driven Modeling (DDM). DDM is based on analyzing data about a system, in particular finding connections between the system state variables (input, internal and output variables) without explicit knowledge of the physical behavior of the system (Solomatine et al., 2008). Therefore, the models carefully generated by data mining tools reflect the real behavior of the system of study, as they are based on the real performance of the system, no matter how complex it is.

There are many significant research studies and experimental works carried-out using various data mining techniques which prove the importance and need of these technologies for oil field applications and water resource management as well (Ahmadi et al., 2017a, b; Li and Li, 2013; Mohan and Ramsundram, 2016; Nourani et al., 2017; Wachowicz, 2002).

A new methodology based on data mining techniques is thus proposed in this research for prediction of WC/SCT in two different types of oil wells discussed previously. The

Fig. 2. Trends of daily BSW production in percentage plotted over time for two real SAOWs a) Well 1, and b) Well 2, both located in Southwest Iran.

main objective is to provide DDMs for prediction and forecasting of WC/SCT to deal with different water production patterns in oil wells. Therefore no assumptions about the conditions of salt production of the investigated wells need to be made by the proposed DDM approach. This research does not consider a modification and/or improvement to the existing methods; rather, it attempts to propose a novel and different approach using the data-driven algorithms for predicting SCT or WC in different conditions of water productions in oil wells.

In the following sections, some required theoretical backgrounds are primarily introduced in Section 2 before proceeding to the proposed approach. After that, Section 3 introduces the proposed data-driven methodologies for predicting and forecasting SCT/WC values in different types of oil wells including SWOWs and SAOWs. The proposed approach will be verified and discussed in Section 4 using data of three real case studies of Iranian oil wells, including one SWOW and two SAOWs. Finally, the significant findings and important remarks that are inferred from this research are listed in Section 5.

2 Theoretical background

Since part of the methodology proposed in this article utilizes special time series data mining techniques including vectorized AR model to forecast the amount of SCT in some kinds of oil wells, the relevant aspects of AR modeling and a brief description of the mentioned algorithm is introduced in the following subsections, before the main body of the proposed approach is introduced.

2.1 Autocorrelation and partial autocorrelation

Autocorrelation and Partial Autocorrelation Functions (ACF and PACF) are the two well-known identification methods of AR models (McLeod and Li, 1983). ACF is defined as the coefficient of correlation between two values

in a time series. Sometimes we may wish to only measure the association between the current value of a time series, y_t and a previous value located at a lag k , y_{t-k} , and filter out the linear influence of the random variables that lie in between (*i.e.*, $y_{t-1}, y_{t-2}, \dots, y_{t-(k-1)}$), which requires a transformation on the time series. Then PACF is obtained by calculating the correlation of the transformed time series.

The PACF is most useful for identifying the order of an AR model (Hamilton and Watts, 1978). Specifically, sample partial autocorrelations that are significantly different from 0 indicate lagged terms of y that are useful predictors of y_t . Graphical approaches to assessing the lag of an AR model include looking at the ACF and PACF values *versus* the lag. In a plot of ACF *versus* the lag, if we see large ACF values and a non-random pattern, then likely the values are serially correlated. In a plot of PACF *versus* the lag, the pattern will usually appear random, but large PACF values at a given lag indicate this value as a possible choice for the order of an AR model.

2.2 Vector AR model

The Vector AR (VAR) model is one of the most successful, flexible, and easy to use models for the analysis of multivariate time series. It is a natural extension of the univariate AR model to dynamic multivariate time series (Lütkepohl, 2006).

A VAR model applies when each variable in the system does not only depend on its own lags, but also the lags of other variables. A general VAR(s) process with white noise can be written as equation (1):

$$y_t = \sum_{j=1}^p \Phi_j y_{t-j} + \epsilon_t, \quad (1)$$

where the error terms follow a vector white noise, *i.e.*, $E(\epsilon_t) = 0$ and y_t is a $k \times 1$ vector of endogenous variables.

Based on the same sample size, the following information criteria and the final prediction error are computed

(Akaike, 1974; Hamilton, 1994; Hannan and Quinn, 1979; Lütkepohl, 2006; Quinn, 1980; Schwarz, 1978):

$$AIC(n) = \ln \det \left(\sum_{\epsilon}^{\sim} (n) \right) + \frac{2}{T} nk^2, \quad (2)$$

$$HQ(n) = \ln \det \left(\sum_{\epsilon}^{\sim} (n) \right) + \frac{2 \ln(\ln(T))}{T} nk^2, \quad (3)$$

$$SC(n) = \ln \det \left(\sum_{\epsilon}^{\sim} (n) \right) + \frac{\ln(T)}{T} nk^2, \quad (4)$$

$$FPE(n) = \left(\frac{T + n^*}{T - n^*} \right)^k \det \left(\sum_{\epsilon}^{\sim} (n) \right), \quad (5)$$

with $\sum_{\epsilon}^{\sim} (n) = T^{-1} \sum_{t=1}^T \hat{\epsilon}_t \hat{\epsilon}_t'$ and n^* is the total number of parameters in each equation and n assigns the lag order. The lag order minimizing each of the above criteria could be the selected lag size for VAR(p) model.

While creating VAR model of the best-determined lag order, it should be noted that not all the coefficients included in the model may have a significant effect on the fitting process. Some of them are negligible and might be removed from the fitted model as they do not provide further improvement in model generation. The t -statistic provides a measure of how extreme a statistical estimate is (Runkel, 2016). The P -value approach involves determining “likely” or “unlikely” by determining the probability of observing a more extreme test statistic in the direction of the alternative hypothesis than the one observed. If the P -value is less than (or equal to) a predefined significance level, α , then the null hypothesis is rejected in favor of the alternative hypothesis, and *vice versa*. In many applications, a P -value less than $\alpha = 0.05$ indicates that the corresponding variable should be considered as significant.

3 Proposed methodology

Two different types of oil wells with variable salt production patterns, including SWOWs and SAOWs, will be studied and modeled in this article, as described in the following statements:

(1) *SWOW*: An innovative methodology for predicting excessive water (salt) production in SWOWs where no specific methods have not been established by other researchers previously is developed in this section. Predicting the amount of water that will be produced aberrantly in substantial volume at future times is considered as the main target for this type of well.

If some thresholds are defined for WC/SCT, the water production data could be easily transformed into some limited number of bins or classes. As negligible water is normally produced during the life of the well, the conditions of increasing water volume (defined as anomalies) could be effectively predicted using a classification method where

two or more class labels are assigned to the input data points. In the simplest case a binary classifier would be used to predict the abnormal conditions *versus* the normal conditions as two different classes. Classification is a method for partitioning data into classes and then attributing data vectors to these classes. The output of a classification model is a class label, rather than a real number like in regression models. Each category of SCT in the classification problem denotes a predefined SCT range or cut-off rather than a single number.

(2) *SAOW*: The aim is, here, to forecast the continuous amount of WC/SCT of the produced fluid over future times for SAOWs. The continuous production of water (or salt concentration) in different volumes provide a suitable framework for using the time series regression analysis methods in order to forecast the WC/SCT of the produced oil. Autoregressive (ARs) models are one of the most powerful and common methods used for forecasting univariate time series (Pandit and Wu, 1983). They use a limited number of past values of the original time series to forecast the variable of interest in the future. It is, therefore, a suitable technique for forecasting the continuous changes in variables over time as in the present case for SAOWs. Due to multiple time series variables that may be involved in the modeling process, a multivariate modification of the AR models, known as Vector AR, is recommended for this purpose in this research. AR methods could be easily found out in McLeod and Zhang (2006).

The complete flowchart of the proposed data-driven methodology for predicting the WC/SCT values in both situations is illustrated in Figure 3. As it is clear from the flowchart, the proposed methodology consists of different parts and processes for each type of well that should be carefully followed by the user. Each part of the presented flowchart is described briefly in the following sections.

3.1 WC/SCT forecasting in SWOWs

In line with the flowchart depicted in Figure 3, the proposed methodology for prediction of SCT/WC in SWOWs using the classification approach is described in the following sections.

3.1.1 Proposed methodology for forecasting WC/SCT in SWOWs: A classification approach

A classification approach is employed to predict the class label of each data point when predicting SCT/WC in SWOWs since they are transformed into several categories or bins prior to creating the model. The required steps proposed to achieve the best classification performance and the most accurate classifier for this purpose are listed below:

(a) Constructing initial dataset

The SCT/WC is considered as the output time series data for the classification problem, while any other available data recorded over time could be regarded as input data. In a daily time scale, the Oil Production Rate (OPR), Cumulative Oil Production (CUMPROD) between two successive days of shutting in the well due to rises in

Fig. 3. Flowchart of the proposed data-driven methodology for forecasting SCT/WC values for the two types of wells considered including SWOW and SAOW.

salt production, and the distance from bottom of production interval of the well to the water-oil-contact (DW) are among the most common variables that are used as input

time series data in this article. To be more clarified about the meaning of CUMPROD, the reader should note that *this is not* the accumulated volume of produced oil from

the starting day of production until the end date; however, it is the cumulative oil volume started to measure from the day when the well is put on production until the day when the well is closed due to significant water production. After closing the well, this parameter is immediately set to zero, until the well is again opened to produce after some days of shut-in. In other words, CUMPROD is equal to zero at time periods when the well is shut-in due to excessive water production and significant flowing pressure drop. As the well is started to flow, the CUMPROD begins to build up again.

(b) Data preprocessing

Handling missing data and replacing them with known or regressed values, removing outliers and vague values from the data, converting SCT data to different numbers of bins/classes, and normalization of individual time series data are among the most important preprocessing techniques in this regard.

(c) Changing data representation

The time series datasets are converted to the conventional non-temporal datasets using the sliding window approach. By this technique, k numbers of successive data points are extracted from each time series using the sliding windows of size k , and then the extracted data points from different time series are arranged in different columns. The m time series of order n will be transformed to a dataset of $n-k$ rows and $m \cdot (k+1)$ columns using the sliding windows of length k . The columns in the new feature space correspond to the lagged variables in the original space.

A grid search approach using repeated Cross Validation (CV) technique is recommended to select the best window size for this purpose. Modeling is performed using different training datasets generated by multiple window sizes and the window length that causes the best classification performance is finally selected as the best window size. CV is employed to avoid over-fitting and also ensure that every example from the original dataset has the same chance of appearing in the training and testing dataset (Kohavi, 1995). The basic form of CV known as k -fold CV will be employed in the present application.

(d) Optimizing classifiers

Different classification algorithms are considered for prediction problem in SWOWs. Each classifier has its own tuning parameters that needs to be optimized. Finding out the best set of tuning parameters for each classifier is reached throughout the grid search technique and repeated CV for resampling training data.

(d1) Partitioning dataset

Prior to modeling, the initial dataset should be partitioned into training and test set to evaluate the model performance by making predictions against the test set.

(d2) Balancing dataset

The training dataset is highly imbalanced because of large differences between the class frequencies. The majority class in the present application belongs to the conditions where nil amounts of salt is produced along with oil in SWOWs. The production of higher amounts of salt occurs very rarely during the production period of the well that will constitute the minority classes. A dataset is said to be imbalanced when the class of interest (minority class) is much fewer than the normal behavior (majority class). The cost of missing a minority class is typically much higher than missing a majority class. Most learning systems are not prepared to cope with imbalanced data and several techniques have been proposed by different authors.

The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. There are a few resampling techniques for this purpose including *random undersampling*, *oversampling* and Synthetic Minority Oversampling Technique (SMOTE) (Mostafizur and Davis, 2013). SMOTE has the merit of avoiding overfitting which occurs when exact replicas of minority instances are added to the main dataset as in the case of oversampling. In addition, loss of useful information which is a disadvantage of random undersampling techniques does not occur with SMOTE. The SMOTE algorithm is therefore recommended as a proper technique for balancing data in this article. The reader is referred to Chawla et al. (2002) for studying the procedure of SMOTE algorithm.

(d3) Averaging classification performance

For each grid point on the dimensional grid search space, estimate the mean performance criterion of the classifier by averaging over all repetitions.

(d4) Finding out best set of tuning parameters

Iterating on all grid points on the search space and evaluating the mean performance measure for each possible hyperparameter, the best set of tuning parameters corresponding to the best performance measure is determined (Ben-Hur and Weston, 2010).

(e) Predicting with test data

Use the well-tuned classifier to predict SCT/WC using the test data and estimate the prediction accuracy of the classifier based on the specific metrics. This is the best classification performance obtained for the given size of sliding window discussed in step c.

(f) Selecting best window size

Iterate on all window sizes so that the most accurate classifier with the best performance criterion is selected at each window size. The optimal window length is estimated by finding out the best classification performance over the window size search grid.

Table 1. List of tuning parameters of different classifiers that are found via grid search and repeated CV in this article.

Classifier	Radial SVM	RF	CT	MLP	LDA	NB
Tuning parameters	Soft-margin constant (C), kernel scaling factor (Sigma [σ])	Number of predictors (m_{try}), number of trees (n_{tree})	–	Hidden layer size	–	–

3.1.2 Classification algorithms for prediction of SCT/WC in SWOWs

There are a variety of well-developed algorithms that could be employed for classification problems. Artificial neural networks and Multi-Layer Perceptron, MLP (Rosenblatt, 1961), Linear Discriminant Analysis, LDA (Raschka, 2014), Classification Tree, CT (Breiman *et al.*, 1984), Random Forest, RF (Breiman, 2001), Support Vector Machine, SVM (Vapnik, 1995) and Naïve Bayes, NB (McCallum and Nigam, 1998) are the common classifiers in this regard. These classification methods that are among the most influential data mining algorithms in the research community (Wu *et al.*, 2008) are to be investigated for predicting WC/SCT of SWOWs using the proposed methodology in this article.

The list of tuning parameters of these classifiers are tabulated in Table 1.

3.1.3 Performance measures of classification

The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives). These four counts constitute a table known as confusion matrix that is often used to describe the performance of a classification model (Sokolova and Lapalme, 2009). In addition, Cohen's Kappa statistic is a very good measure that can handle very well both multi-class and imbalanced class problems (Cohen, 1960). Cohen's Kappa is always less than or equal to 1. Values of 1 indicate perfect agreement, while a value of zero would indicate a lack of agreement (Landis and Koch, 1977).

3.2 WC/SCT forecasting in SAOWs

According to the flowchart illustrated in Figure 3, the proposed methodology for forecasting SCT/WC in SAOWs using the regression approach is designated in the subsequent sections. The proposed approach uses a vectorized AR model to conduct the forecasting process for the current problem. The related aspects of AR modeling and a brief description of the above-mentioned algorithm was described previously in Section 2.

3.2.1 Proposed methodology for forecasting WC/SCT in SAOWs: a regression approach

The steps that are recommended in this article for forecasting SCT/WC values over time in SAOWs using VAR modeling is as follows:

a) Dataset preparation

The whole dataset including different time series variables related to SCT/WC production in oil wells is constructed.

b) Data preprocessing

Handling missing data and replacing them with known or regressed values, and data normalization are the most required preprocessing techniques for this purpose.

c) Data partitioning

Split the dataset into training and testing datasets initially.

d) VAR selection

Select the best lag size, p , based on the highest values of SC or AIC measures, preferably. The best lag order may be also found using a grid search approach by running VAR models of different lag orders from 1 to the maximum value determined by different VAR selection criteria. The lag order that yields the lowest forecasting error for the test data will be used as the best lag size in creating the VAR model. The Mean Absolute Error (MAE), defined as the mean of the absolute values of the individual prediction errors over all instances in the test set, can be used as the forecasting performance criterion in this regard.

e) Evaluating VAR model against test data

Forecast the SCT/WC for some future days using the best VAR model. Compare the forecasted values with the target data and estimate the forecasting performance (MAE, RMSE, etc.) as the final measure of regression accuracy.

4 Results and discussion

To validate the approaches proposed in this research for predicting the amount of SCT in different kinds of oil wells, three real case studies including one SWOW (Well A) and two SAOWs (Wells B and C) from two different oil reservoirs both located in Southwest Iran are implemented in the following sections using the methodologies proposed in Section 3. The modeling procedures are performed using MATLAB and *R* statistical languages in this article.

The study variables used in this article include WC/SCT as the target data and some other related parameters

Fig. 4. Plots of a) fitting WC *versus* SCT data, and b) their corresponding residuals for all BSW data of Well A including outliers.

like OPR, CUMPROD and DW (if available) as the input data for several consecutive days of production.

Some entries in the gathered BSW tables are missing for either the amount of SCT or WC percentage. A suitable preprocessing technique is therefore needed to handle the missing data values prior to modeling. One of the common methodologies used for this purpose is the replacement of missing data with some interpolated values. A simple mathematical correlation, if accessible, could be found to relate the fraction of water produced along with the oil (*i.e.* surface WC) to the amount of salt concentration in the produced fluid (*i.e.* the SCT). It's possible to use such a fitted correlation to convert the SCT and WC values to each other. Therefore a scatter plot of WC percentage *versus* SCT values is generated for the rows with complete WC and SCT entries to extract the correlation between them, as illustrated in Figure 4a. This figure also displays the best fitted line passing through the data points. The calculated residuals between the actual and predicted WC values while considering the whole dataset in the fitting process are plotted in Figure 4b as well. Some points shown in small red crosses on the plots seem to be outliers, as they locate relatively far distances from the mean of data. One of the most common ways of visually identifying outliers is the

box-plot tool (Seo, 2006) that is used in this article. To improve performance of the fitting process, the outliers have to be excluded from the training dataset prior to regression. The plots of fitting WC *versus* SCT values and their corresponding residuals while disregarding the outliers from data are displayed in Figures 5a and 5b, respectively. The equation of the best-fitted line and its corresponding regression measures are provided in Table 2. Using this equation, one can handle missing values of SCT or WC data in the data set. The linear equation fitted in this article is based on the data gathered from a number of oil fields, with near equal total formation water salinity, including the two reservoirs where the three real case studies Wells A, B and C are located in.

4.1 Forecasting SCT in SWOW using different classifiers

The methodology proposed in Section 3.1.1 is verified using the real data of a SWOW (labeled as Well A) produced from an oil reservoir connected to a weak aquifer in South-west Iran. The endogenous variables for this well include WC/SCT as the input data and OPR, CUMPROD and DW as the output data for 3349 successive days.

Fig. 5. Plots of a) fitting WC *versus* SCT data, and b) their corresponding residuals for training data of Well A excluding outliers.

Table 2. Information of best-fitted line of WC data *versus* SCT values

Equation form	$WC = p_1 \times (SCT) + p_2$
Coefficients (with 95% confidence bounds)	$P_1 = 0.0005487 \text{ (0.0005424, 0.0005551)}$ $P_2 = 0.1835 \text{ (0.0775, 0.2895)}$
Regression error	RMSE = 0.6941

The plots of study variables including SCT (in mg/L), DW (in m), OPR (in STB/day) and CUMPROD (in STB) for Well A are shown in Figures 6a through 6d, respectively. As shown in Figure 6a, the SCT values are below the permissible threshold for very large portions of the data; it rises above the threshold, however, only for a few days as shown by limited peaks on the graph. The methodology presented in Section 3.1.1 for prediction of SCT/WC in SWOWs is thoroughly examined for the present case study.

According to the regulatory bases, the cut-offs displayed in Table 3 were chosen for discretization of SCT values in this article.

Next, the class imbalance problem as depicted from the plot of Figure 6a is alleviated using the SMOTE algorithm before proceeding to the modeling phase. The number of

instances of different classes both before and after balancing data is shown in Table 4.

Six different classifiers, as discussed in Section 3.1.2, are investigated using the proposed methodology. The validation method and its related parameters used in modeling classifiers are displayed in Table 5.

Several R packages including CARET (Kuhn *et al.*, 2016), PARTY (Hothorn *et al.*, 2006), Random Forest (Liaw and Wiener, 2002), klaR (Weihs *et al.*, 2005), kernlab (Karatzoglou *et al.*, 2004), RSNNS (Bergmeir and Benitez, 2012) and R.matlab (Bengtsson, 2016) are employed in this section to execute different classifiers.

The results achieved for the present case study including the smallest, the mean and the largest Kappa coefficients as well as the best size of sliding window and the best tuning

Fig. 6. Original time series plots for Well A; there are plots of four variables ordered in time (in a daily scale) including a) SCT in mg/L, b) DW in m, c) OPR in STB/day, and d) CUMPROD in STB.

Table 3. The extreme values and cut-offs for binning SCT data of Well A.

Min. SCT	Max. SCT	Cut-off 1	Cut-off 2	Cut-off 3
19	400	43	80	300

parameters of each classifier (corresponding to the largest Kappa coefficient values) are summarized in Table 6.

The plots of Kappa coefficient and classification accuracy versus the size of sliding window is also illustrated in Figures 7a through 7f for RF, CT, MLP, Radial SVM,

LDA and NB classifiers, respectively. To make a better comparison between all classifiers, the plots of maximum Kappa coefficients and the best sizes of sliding windows are depicted in Figure 8 for all classification algorithms at the same time. The left vertical axis of the plot shows the best size of sliding window of different classifiers marked by small triangles, while the right axis displays the largest Kappa coefficients marked by circles. According to these plots, the ranking of classifiers in terms of descending Kappa coefficient is as follows: RF, CT, MLP, Radial SVM, NB and LDA. In other words, the largest and the most plausible Kappa coefficients are achieved by RF, CT, MLP and Radial SVM algorithms, respectively. The

Table 4. The number of members of different classes before and after balancing data.

Conditions	Number of examples in class 1 (Nil SCT or normal conditions)	Number of examples in class 2	Number of examples in class 3	Number of examples in class 4	Number of examples in classes 2, 3 and 4 (abnormal conditions)	Total number of examples	Ratio of examples of abnormal to normal conditions
Imbalanced data	3336	4	6	3	13	3349	0.0039
Balanced data	3336	1672	1674	1674	5020	8356	0.600

Table 5. Methods and parameters used in classification procedure using the proposed methodology.

Validation technique	CV technique	Number of CV repetitions	Range of sliding window size	Classification performance measure
Repeated CV	k -fold ($k = 5$)	20	0–15	Kappa coefficient

Table 6. Results of SCT prediction for SWOW A using six different classifier and the proposed methodology.

Classifier	Smallest kappa coefficient	Mean kappa coefficient	Largest kappa coefficient	Best sliding window size (k)	Best tuning parameters
RF	0.996	0.996	0.997	12	$m_{\text{try}} = 7$; $n_{\text{tree}} = 84$
CT	0.750	0.929	0.992	1	–
MLP	0.901	0.954	0.980	14	Size = 28
Radial SVM	0.915	0.939	0.955	10	Sigma (σ) = 0.1367; $C = 65536$
LDA	0.167	0.465	0.616	15	–
NB	0.432	0.484	0.623	0	–

other classifiers including NB and LDA have yielded relatively low values of Kappa coefficient for classifying SCT in the present problem.

Among the best classifiers, the shortest size of sliding window belongs to CT which is equal to 1. This means that CT uses the input variables at the current time step and those recorded at the most recent time step (*i.e.* 6 input variables totally). In addition, RF uses only seven input variables at best condition for predicting SCT according to the entries of best tuning parameter (m_{try}) of Table 6, although the best sliding window size for this algorithm is relatively large ($k = 12$). Based on the smallest Kappa coefficient values provided in Table 6, the RF classifier also achieves a very good performance using the shorter sliding windows. For windows of sizes 0 and 1, for example, the Kappa coefficients of the best model instances of RF classifier are both equal to about 0.996 which is a little bit larger than that of the best model instance of CT (*i.e.*, 0.992). Alternatively, CT and RF have the lowest complexity in terms of the number of input variables compared to MLP and Radial SVM as they utilize more numbers of previous

values for providing the best Kappa coefficients. The NB algorithm, however, has the shortest size of sliding window (equal to 0) meaning that it exploits only the input variables at the current time step regardless of the previous time steps. Therefore, NB performance using the window size of 0 is comparatively lower than the best performances provided by the above classifiers and increasing the length of sliding window just degrades the Kappa coefficient of NB. For LDA, we encounter a lazy classifier as it could not predict the classes correctly even with the best window size of 15. Trivially, the prediction performance of LDA is very low using the windows of small sizes.

The confusion matrix plots of the predicted values *versus* the targets using the best model of each classifier (RF, CT, MLP, SVM, LDA and NB, respectively) are illustrated in Figures 9 and 10 for the whole dataset and the test set respectively. According to Figures 9a through 9d, for the first four classification algorithms, the classes C_2 through C_4 corresponding to the SCT values greater than the maximum allowable content for SWOWs are predicted with 100% accuracy compared to the real targets. If binary

Fig. 7. Plots of classification performance measures (including Kappa coefficient and accuracy values) *versus* sliding window size for different classifiers, a) RF, b) CT, c) MLP, d) Radial SVM, e) LDA, and f) NB.

classification with two class labels including “under-threshold” and “over-threshold” SCT is expected, then the NB algorithm also provides a perfect performance in

predicting the real classes of data points belonging to the over-threshold SCT. Acceptable results are, however, not obtained using the LDA classifier for the training and test datasets for the present case study. Similar conclusions could be inferred from Figures 10a through 10f about the performance of different classifiers when predicting the test dataset only.

Fig. 8. Plots of best window size (k) and maximum Kappa coefficient for six classifiers in this article.

4.2 Forecasting SCT in SAOWS using VAR models

The methodology proposed in Section 3.2.1 is now verified using the real data of two SAOWs (named as Wells B and C) produced from a water-drive oil reservoir in Southwest Iran. The two wells have nearly the same production conditions and are completed with similar downhole structures in the same formation. The endogenous variables available for the two wells include SCT, OPR and CUMPROD data for 395 successive days while DW data was not available for them.

Fig. 9. Confusion matrix plots of best model instances of each classifier used in predicting SCT classes for the *whole* balanced dataset of SWOW A: a) RF, b) CT, c) MLP, d) Radial SVM, e) LDA, and f) NB.

Figures 11 and 12 illustrate the input and output time series data (each including SCT, OPR and CUMPROD, respectively) plotted in a daily time scale for Wells B and C, individually. Looking at the history plots of SCT in Figures 11a and 12a, the cyclic and periodic patterns with

non-monotonic trends are easily observed during the time of study.

Normalization is first applied to the whole dataset so that the values in each time series are transformed between 0 and 1. Prior to forecasting, a limited number of time data

Fig. 10. Confusion matrix plots of best model instances of each classifier used in predicting SCT classes for the *testing* dataset of SWOW A: a) RF, b) CT, c) MLP, d) Radial SVM, e) LDA, and f) NB.

points (here the last 30 points) is considered as test data and are excluded from the training phase.

For each well, VAR is conducted on multiple time series based on the proposed methodology to forecast the SCT over future times. This is implemented using the VARS package in the *R* environment (Pfaff, 2008). The values of different criteria including AIC, HQ, SC and FPE for

different lag orders from 1 to 30 are plotted in Figures 13 and 14 for Wells B and C, respectively. The lag orders that minimize each criterion is chosen as the selected lag size for the corresponding criteria as displayed in Table 7 for both Wells B and C. The best lag size is selected using the grid search approach by running a variety of VAR models of different lag orders from 1 to the maximum value selected

Fig. 11. Time series plots of Well B for SCT forecasting problem, a) SCT in mg/L, b) OPR in STB/day, and c) CUMPROD in STB.

by the four criteria individually for each well. The plots of MAE measures of test data for a number of VAR models with different lag orders for Wells B and C are illustrated in [Figures 15 and 16](#), respectively. For Well B, the lowest possible MAE measure of 0.062 happens for a lag order of $p = 4$. Considering Well C, on the other hand, the lowest possible MAE measure of 0.0975 happens for the lag order of $p = 1$. The final VAR(4) and VAR(1) models are created using the normalized training time series data without any special transformations for Wells B and C, respectively. The corresponding equations of VAR(4) and VAR(1) models are given by equations (6) and (7):

$$\begin{aligned} \text{SCT} = & \sum_{i=1}^2 A_i \text{SCT}.li + \sum_{i=1}^2 B_i \text{OPR}.li \\ & + \sum_{i=1}^2 C_i \text{CUMPROD}.li + \text{const}, \end{aligned} \quad (6)$$

$$\text{SCT} = A_1 \text{SCT}.l1 + B_1 \text{OPR}.l1 + C_1 \text{CUMPROD}.l1 + \text{const}, \quad (7)$$

where A_i , B_i and C_i are the coefficients for SCT, OPR and CUMPROD variables in the model, respectively. In addition, li indices of each variable indicate the i most previous time steps of the corresponding variable at time $t-i$. As discussed in [Section 2.2](#), the importance of coefficients

in the created VAR model should be evaluated using the t -statistic and the corresponding P -values. In the present application, a P -value less than $\alpha = 0.05$ indicates that the corresponding variable should be considered as significant. The coefficient estimates and their standard errors of VAR models along with their corresponding t -statistic and P -values for SCT data are therefore summarized in [Tables 8 and 9](#) for Wells B and C, respectively. The significant variables having a relatively small P -value are indicated by a star sign in the last column of [Tables 8 and 9](#). Other variables not indicated as significant in the Table will not be included in the final model. The residual standard error and coefficient of determination (R -squared) on the whole training data set is also shown in the lowest part of [Tables 8 and 9](#).

Diagrams of fit and residuals for SCT data over time for Wells B and C are shown in [Figures 17a and 17b](#) and [Figures 18a and 18b](#), respectively. The plots demonstrate very good accuracy in forecasting the SCT data of the training dataset. In addition, the ACF and PACF plots of residuals *versus* lag orders are displayed in [Figures 17c and 17d](#) and [Figures 18c and 18d](#), respectively, for Wells B and C. The autocorrelation plots also verify that the fitting residuals are not correlated in any lag order for both wells.

The forecasted normalized SCT data with 95% confidence intervals for 30 days ahead of the training dataset (*i.e.* for the test dataset) using the generated VAR models is depicted in [Figure 19](#), for Well B. To make a better

Fig. 12. Time series plots of Well C for SCT forecasting problem, a) SCT in mg/L, b) OPR in STB/day, and c) CUMPROD in STB.

Fig. 13. Plots of different selection criteria *versus* lag size for VAR(p) model for forecasting SCT data of Well B.

Fig. 14. Plots of different selection criteria *versus* lag size for VAR(p) model for forecasting SCT data of Well C.

comparison, the true and forecasted normalized salt concentration data for Well B are plotted on the same graph as displayed in Figure 20. Figure 20a shows this comparison only for 30 test data points, while the entire dataset comparison is displayed in Figure 20b. For each graph, the relative errors between the target and forecasted data points are also displayed on the corresponding plot. Such graphs, as created for Well B, are illustrated in Figures 21

Table 7. Best lag size of VAR(p) model, based on different statistical criteria, for forecasting SCT data of Wells B and C.

Well tag	AIC (n)	HQ (n)	SC (n)	FPE (n)
B	28	8	2	28
C	27	7	7	27

Fig. 15. Plot of MAE values of test data *versus* lag size to select the best lag order of VAR model for Well B.

Fig. 16. Plot of MAE values of test data *versus* lag size to select the best lag order of VAR model for Well C.

Table 8. Coefficient information of VAR(4) model for SCT forecasting for Well B.

Variable	Estimate of coefficient	Standard deviation of coefficient	<i>t</i> -statistic	<i>P</i> -value	Importance
SCT.11	1.635	0.055	29.996	<2.00E-16	*
OPR.11	−0.024	0.054	−0.441	0.659	—
CUMPROD.11	0.229	0.072	3.188	0.002	*
SCT.12	−0.656	0.106	−6.178	0.000	*
OPR.12	0.008	0.074	0.113	0.910	—
CUMPROD.12	−0.444	0.155	−2.854	0.005	*
SCT.13	0.182	0.103	1.763	0.079	—
OPR.13	−0.009	0.073	−0.117	0.907	—
CUMPROD.13	0.070	0.159	0.442	0.659	—
SCT.14	−0.188	0.052	−3.615	0.000	*
OPR.14	0.000	0.054	0.008	0.994	—
CUMPROD.14	0.135	0.075	1.804	0.072	—
const	0.016	0.004	3.492	0.001	*
Residual standard error		0.018	—	—	—
<i>R</i> -squared		0.994	—	—	—

The asterisk symbol in the last column shows the significance of the corresponding model parameter.

Table 9. Coefficient information of VAR(1) model for SCT forecasting problem for Well C.

Variable	Estimate of coefficient	Standard deviation of coefficient	<i>t</i> -statistic	<i>P</i> -value	Importance
SCT.11	0.952	0.016	61.199	<2.00E-16	*
OPR.11	0.023	0.010	2.354	0.019	*
CUMPROD.11	−0.059	0.023	−2.546	0.011	*
const	0.017	0.009	1.978	0.049	*
Residual standard error		0.0478	—	—	—
<i>R</i> -squared		0.919	—	—	—

The asterisk symbol in the last column shows the significance of the corresponding model parameter.

and 22 for Well C, respectively. As the graphs show for both wells, except for some limited data points, the forecasted data values are in good agreement and well coincide with

the true data for a large portion of both training and testing data points. The estimated values of relative errors, as shown in Figures 20 and 22, verify this conclusion for both

Fig. 17. Plots of SCT forecasting for Well B using VAR algorithm including a) diagram of fit for SCT, b) diagram of residuals for SCT, c) ACF of residuals, and d) PACF of residuals.

Fig. 18. Plots of SCT forecasting for Well C using VAR algorithm including a) diagram of fit for SCT, b) diagram of residuals for SCT, c) ACF of residuals, and d) PACF of residuals.

Fig. 19. Forecasting results of normalized SCT of Well B for 30 days ahead of the training dataset using VAR(4) model; the plot shows the forecasted values (shown in green circles) along with the upper and lower bounds (shown in blue squares and orange triangles, respectively), corresponding to 95% confidence intervals.

Fig. 20. Investigating through the forecasting performance of VAR model for Well B, a) normalized original and forecasted *test* data points along with their relative errors plotted on the same graph, b) normalized original and forecasted data points for the *entire* data set including both training and testing data along with their relative errors plotted on the same graph. On both plots, the vertical and horizontal axes show the normalized SCT data and time order (in days), respectively. The original and forecasted data points are shown by small blue circles and orange triangles, respectively, while the relative errors are depicted by small gray diamonds.

Fig. 21. Forecasting results of normalized SCT of Well C for 30 days ahead of the training dataset using VAR(1) model; the plot shows the forecasted values (shown in green circles) along with the upper and lower bounds (shown in blue squares and orange triangles, respectively), corresponding to 95% confidence intervals.

Fig. 22. Investigating through the forecasting performance of VAR model for Well C, a) normalized original and forecasted *test* data points along with their relative errors plotted on the same graph, b) normalized original and forecasted data points for the *entire* data set including both training and testing data along with their relative errors plotted on the same graph. On both plots, the vertical and horizontal axes show the normalized SCT data and time order (in days), respectively. The original and forecasted data points are shown by small blue circles and orange triangles, respectively, while the relative errors are depicted by small gray diamonds.

Wells B and C, as well. The VAR models are able to forecast the real patterns of salt production for the test data with estimated MAE values equal to 0.062 for Well B and 0.0975 for Well C. These values show quite good accordance between the real and forecasted SCT data for 30 last data points using the well-tuned VAR models.

5 Conclusion

The main contribution of this article is to propose a novel data-driven approach for forecasting the salt content (or equivalently the water-cut) of producing oil wells at different conditions. These conditions are divided into two different categories including continuous water production (for salty oil wells) and discontinuous water production (for sweet oil wells). The approaches presented in this article for the prediction of SCT/WC in these two conditions are different in methodology and algorithms. A classification approach is proposed for predicting the conditions of excessive salt content production (larger than a pre-specified threshold) for sweet oil wells, while a regression-based methodology is presented for forecasting the amount of SCT over near future times for salty oil wells. The conclusions inferred from the present study are summarized in the following statements.

1. Prediction of irregular and continuous production of SCT/WC over times in SWOWs and SAOWs is investigated using DDM in this article.
2. This article provides a new methodology for forecasting WC/SCT production using multiple variables recorded in time including OPR, CUMPROD and DW. Any other related temporal variables that become available could be easily incorporated into the model just by adding new dimensions to the feature space to improve the model accuracy.
3. A classification approach is proposed in this article for SCT prediction in SWOWs where significant salt production occurs rarely during well production period.
4. Six different classification algorithms including RF, CT, MLP, Radial SVM, NB and LDA are examined using the proposed methodology for SCT production in SWOWs. According to the results of a real case study, the RF, CT, MLP and Radial SVM can provide the best performance measures (the largest Kappa coefficients), respectively.
5. The well-tuned CT and RF classification algorithms created for the present case study utilize relatively lower numbers of input variables compared to MLP and Radial SVM.
6. A VAR model approach is employed in this article to forecast SCT/WC in SAOWs where the well continuously produces brine throughout its production period. The proposed VAR modeling approach is verified using data of two real SAOWs. The wells show cyclic and periodic patterns of varying trends for water production. A VAR(4) and a VAR(1) model are created for the two case studies based on the lowest values of MAE measures achieved for forecasting

SCT data for 30 days ahead of the training data. As the results of the present case studies show, the well-tuned VAR models generated using the proposed methodology in this research could provide reliable and acceptable results with reasonable accuracy in forecasting the SCT values for the near future days.

7. The DDM approach proposed in this article overcomes the main limitation of the empirical methods of WC prediction in using the simplifying assumptions and handling a limited number of independent variables in the prediction models. Besides, the VAR modeling approach proposed in this study could effectively deal with and forecast the cyclic patterns and non-monotonic trends of SCT data.

References

- Ahmadi R., Aminshahidy B., Shahrabi J. (2017a) Well-testing model identification using time-series shapelets, *J. Pet. Sci. Eng.* **149**, 292–305. doi: [10.1016/j.petrol.2016.09.044](https://doi.org/10.1016/j.petrol.2016.09.044).
- Ahmadi R., Shahrabi J., Aminshahidy B. (2017b) Automatic well-testing model diagnosis and parameters estimation using artificial neural networks and design of experiments, *J. Petrol. Explor. Prod. Technol.* **7**, 759–783. doi: [10.1007/s13202-016-0293-z](https://doi.org/10.1007/s13202-016-0293-z).
- Akaike H. (1974) A new look at the statistical model identification, *IEEE Trans. Automat. Contr. AC* **19**, 716–723.
- Bengtsson H. (2016) *R.matlab: Read and Write MAT files and call MATLAB from within R*. R package version 3.6.1. <https://CRAN.R-project.org/package=R.matlab>. Accessed 31 May 2018.
- Ben-Hur A., Weston J. (2010) A user's guide to support vector machines, in: O. Carugo, F. Eisenhaber (eds), *Data Mining Techniques for the Life Sciences. Methods in Molecular Biology (Methods and Protocols)*, Humana Press, Vol. 609, doi: [10.1007/978-1-60327-241-4_13](https://doi.org/10.1007/978-1-60327-241-4_13).
- Bergmeir C., Benitez J.M. (2012) Neural networks in R using the Stuttgart neural network simulator: RSNNS, *J. Stat. Softw.* **46**, 1–26. doi: [10.18637/jss.v046.i07](https://doi.org/10.18637/jss.v046.i07).
- Breiman L. (2001) Random forests, *Mach. Learn.* **45**, 5–32. doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) *Classification and regression trees*, Chapman & Hall/CRC, United States.
- Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. (2002) SMOTE: synthetic minority over-sampling technique, *J. Artif. Intell. Res.* **16**, 321–357. doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953).
- Clark C.E., Veil J.A. (2009) *Produced water volumes and management practices in the United States*, Argonne Natl. Lab. Rep., United States, pp. 1–64. doi: [10.2172/1007397](https://doi.org/10.2172/1007397).
- Cohen J. (1960) A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.* **20**, 37–46.
- Ershaghi I., Omorigie O. (1978) A method for extrapolation of cut vs recovery curves, *Soc. Pet. Eng.* **30**, 203–204. doi: [10.2118/6977-PA](https://doi.org/10.2118/6977-PA).
- Hamilton D.C., Watts D.G. (1978) Interpreting partial autocorrelation functions of seasonal time series models, *Biometrika* **65**, 135–140. doi: [10.2307/2335288](https://doi.org/10.2307/2335288).
- Hamilton J. (1994) *Time series analysis*, Princeton University Press, Princeton.

- Hannan E.J., Quinn B.G. (1979) The determination of the order of an autoregression, *J. Royal Stat. Soc.* **B41**, 190–195.
- Hothorn T., Hornik K., Achim Z. (2006) Unbiased recursive partitioning: a conditional inference framework, *J. Comput. Graph. Stat.* **15**, 651–674. doi: [10.1198/106186006X133933](https://doi.org/10.1198/106186006X133933).
- Karatzoglou A., Smola A., Hornik K., Zeileis A. (2004) kernlab – An S4 Package for Kernel Methods in R, *J. Stat. Softw.* **69**, 721–729. doi: [10.18637/jss.v011.i09](https://doi.org/10.18637/jss.v011.i09).
- Kohavi R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2*, Morgan Kaufmann Publishers Inc, San Francisco, pp. 1137–1143.
- Kuhn M., Wing J., Weston S., Williams A., Keefer C., Engelhardt A., Cooper T., Mayer Z., Kenkel B., The R Core Team, Benesty M., Lescarbeau R., Ziem A., Scrucca L., Tang Y., Candan C., Hunt T. (2016) *Caret: classification and regression training*, R package version 6.0-73. <https://CRAN.R-project.org/package=caret>. Accessed 31 May 2018.
- Landis J.R., Koch G.G. (1977) The measurement of observer agreement for categorical data, *Biometrics* **33**, 159–174. doi: [10.2307/2529310](https://doi.org/10.2307/2529310).
- Lawal K.A., Utin E., Langaas K. (2007) A didactic analysis of water cut trend during exponential oil-decline, *Nigeria Annual International Conference and Exhibition*, 6–8 August, Abuja, Nigeria, Society of Petroleum Engineers. doi: [10.2118/111920-MS](https://doi.org/10.2118/111920-MS).
- Li K., Ren X., Li L., Fan X. (2011) A new model for predicting water cut in oil reservoirs, *SPE EUROPEC/EAGE Annual Conference and Exhibition*, 23–26 May, Vienna, Austria, Society of Petroleum Engineers. doi: [10.2118/143481-MS](https://doi.org/10.2118/143481-MS).
- Li X., Li H. (2013) A new method of identification of complex lithologies and reservoirs: task-driven data mining, *J. Pet. Sci. Eng.* **109**, 241–249. doi: [10.1016/j.petrol.2013.08.049](https://doi.org/10.1016/j.petrol.2013.08.049).
- Liaw A., Wiener M. (2002) Classification and regression by random forest, *R News* **2**, 18–22.
- Liu P., Mu Z., Wang W., Liu P., Hao M., Liu J. (2017) A new combined solution model to predict water cut in water flooding hydrocarbon reservoirs, *Int. J. Hydrogen Energy* **42**, 18685–18690. doi: [10.1016/j.ijhydene.2017.04.166](https://doi.org/10.1016/j.ijhydene.2017.04.166).
- Lo K.K., Warner H.R., Johnson J.B. (1990) A study of the post-breakthrough characteristics of waterflood, *SPE California Regional Meeting*, 4–6 April, Ventura, California, Society of Petroleum Engineers. doi: [10.2118/20064-MS](https://doi.org/10.2118/20064-MS).
- Lütkepohl H. (2006) *New introduction to multiple time series analysis*, Springer, New York.
- McCallum A., Nigam K. (1998) A comparison of event models for Naive Bayes text classification. *Proc AAAI/ICML-98 Workshop on Learning for Text Categorization*, AAAI Press, Madison, WI, pp. 41–48.
- McLeod A.I., Li W.K. (1983) Diagnostic checking ARMA time series models using squared-residual autocorrelations, *J. Time Ser. Anal.* **4**, 269–273. doi: [10.1111/j.1467-9892.1983.tb00373.x](https://doi.org/10.1111/j.1467-9892.1983.tb00373.x).
- McLeod A.I., Zhang Y. (2006) Partial autocorrelation parameterization for subset autoregression, *J. Time Ser. Anal.* **27**, 599–612. doi: [10.1111/j.1467-9892.2006.00481.x](https://doi.org/10.1111/j.1467-9892.2006.00481.x).
- Mohan S., Ramsundram N. (2016) Predictive temporal data-mining approach for evolving knowledge based reservoir operation rules, *Water Resour. Manage.* **30**, 3315–3330. doi: [10.1007/s11269-016-1351-5](https://doi.org/10.1007/s11269-016-1351-5).
- Mostafizur M.R., Davis D.N. (2013) Addressing the class imbalance problem in medical datasets, *Int. J. Mach. Learn. Comput.* **3**, 224–228. doi: [10.7763/IJMLC.2013.V3.307](https://doi.org/10.7763/IJMLC.2013.V3.307).
- Nasiri M., Jafari I., Parniankhoy B. (2017) Oil and gas produced water management: a review of treatment technologies, challenges, and opportunities, *Chem. Eng. Commun.* **204**, 990–1005. doi: [10.1080/00986445.2017.1330747](https://doi.org/10.1080/00986445.2017.1330747).
- Nourani V., Sattari M.T., Molajou A. (2017) Threshold-based hybrid data mining method for long-term maximum precipitation forecasting, *Water Resour. Manage.* **31**, 2645–2658. doi: [10.1007/s11269-017-1649-y](https://doi.org/10.1007/s11269-017-1649-y).
- Pandit S.M., Wu S.M. (1983) *Time series and system analysis with applications*, John Wiley & Sons, New York, United States.
- Petrowiki (2018) *Produced oilfield water*. https://petrowiki.org/Produced_oilfield_water. Accessed 22 Jan 2019.
- Pfaff B. (2008) VAR, SVAR and SVEC Models: Implementation Within R Package vars, *J. Stat. Softw.* **27**, 1–32. <http://hdl.handle.net/10.18637/jss.v027.i04>.
- Quinn B. (1980) Order determination for a multivariate autoregression, *J. Royal Stat. Soc.* **B42**, 182–185.
- Raschka S. (2014) *Linear Discriminant Analysis – Bit by Bit*. http://sebastianraschka.com/Articles/2014_python_lda.html#principal-component-analysis-vs-linear-discriminant-analysis. Accessed 31 May 2017.
- Rosenblatt F.X. (1961) *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*, Spartan Books, Washington.
- Runkel P. (2016) *What are T values and P values in statistics? The Minitab blog*, Minitab Blog Editor, United States, <http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-t-values-and-p-values-in-statistics>. Accessed 31 May 2018.
- Schwarz G. (1978) Estimating the dimension of a model, *Ann. Stat.* **6**, 461–464. doi: [10.1214/aos/1176344136](https://doi.org/10.1214/aos/1176344136).
- Seo S. (2006) *A review and comparison of methods for detecting outliers in univariate data sets*, Dissertation, University of Pittsburgh, Pennsylvania.
- Sitorus J.H.H., Sofyan A., Abdulfatah M.Y. (2006) Developing a fractional flow curve from historic production to predict performance of new horizontal wells, Bekasap Field, Indonesia, *SPE Asia Pacific Oil & Gas Conference and Exhibition*, 11–13 September, Adelaide, Australia, Society of Petroleum Engineers. doi: [10.2118/101144-MS](https://doi.org/10.2118/101144-MS).
- SLB (2018) *Water management for oil and gas*. <https://www.slb.com/services/additional/water/oil.aspx>. Accessed 30 May 2018.
- Sokolova M., Lapalme G. (2009) A systematic analysis of performance measures for classification tasks, *Inf. Process. Manage.* **45**, 427–437. doi: [10.1016/j.ipm.2009.03.002](https://doi.org/10.1016/j.ipm.2009.03.002).
- Solomatine D., See L.M., Abraart R.J. (2008) Data-driven modeling: Concepts, approaches and experiences, in: Abrahart R.J., See L.M., Solomatine D.P. (eds), *Practical Hydroinformatics*. Water Science and Technology Library 68. Springer, Berlin, Heidelberg.
- Vapnik V. (1995) *The nature of statistical learning theory*, Springer, New York.
- Wachowicz M. (2002) Uncovering spatio-temporal patterns in environmental data, *Water Resour. Manage.* **16**, 469–487. doi: [10.1023/A:1022259531710](https://doi.org/10.1023/A:1022259531710).
- Weih S., Ligges U., Luebke K., Raabe N. (2005) klaR analyzing German business cycles, in: Baier D., Decker R., Schmidt-Thieme L. (eds), *Data analysis and decision support*, Springer-Verlag, Berlin, pp. 335–343.
- Wu X., Kumar V., Quinlan J.R., Ghosh J., Yang Q., Motoda H., McLachlan G.J., Ng A., Liu B., Yu P.S., Zhou Z.H., Steinbach M., Hand D.J., Steinberg D. (2008) Top 10 algorithms in data mining, *Knowl. Inf. Syst.* **14**, 1–37. doi: [10.1007/s10115-007-0114-2](https://doi.org/10.1007/s10115-007-0114-2).