Lead author
Minou Rabiei

A NOVEL APPROACH IN EXTRACTING PREDICTIVE INFORMATION FROM WATER-OIL RATIO FOR ENHANCED WATER PRODUCTION MECHANISM DIAGNOSIS

M. Rabiei[1], R. Gupta[1], Y.P. Cheong[2] and G.A. Sanchez Soto[2]
[1]Department of Mathematics and Statistics
Curtin University of Technology
Kent St, Bentley
Perth WA 6102
[2]CSIRO, Earth Science and Resource Engineering
26 Dick Perry Ave
Kensington WA 6151
Minou.rabiei@postgrad.curtin.edu.au
R.gupta@curtin.edu.au
Yawpeng.cheong@csiro.au
Gerardo.sanchezsoto@csiro.au

ABSTRACT

Despite the advances in water shutoff technologies, the lack of an efficient diagnostic technique to identify excess water production mechanisms in oil wells is preventing these technologies being applied to deliver the desired results, which costs oil companies a lot of time and money.

This paper presents a novel integrated approach for diagnosing water production mechanisms by extracting hidden predictive information from water-oil ratio (WOR) graphs and integrating it with static reservoir parameters. Two common types of excess water production mechanism (coning and channelling) were simulated where a wide range of cases were generated by varying a number of reservoir parameters. Plots of WOR against oil recovery factor were used to extract the key features of the WOR data. Tree-based ensemble classifiers were then applied to integrate these features with the reservoir parameters and build classification models for predicting the water production mechanism.

Our results show high rates of prediction accuracy for the range of WOR variables and reservoir parameters explored, which demonstrate the efficiency of the proposed ensemble classifiers. Proactive water control procedures based on proper diagnosis obtained by the proposed technique would greatly optimise oil productivity and reduce the environmental impacts of the unwanted water.

KEYWORDS

WOR plots, excess water production, channelling, coning, classification, random forests, ensemble techniques.

INTRODUCTION

In recent years, unconventional mathematical techniques and soft computing methodologies have gained more and more popularity in the oil and gas industry. The complex nature of the oil fields combined with staggering volume and diversity of data and resulting uncertainties calls for more sophisticated techniques to integrate various types of data, quantify uncertainties, identify hidden patterns and extract useful information. Data mining is one of the promising methodologies that can offer great benefits to the oil industry by extracting hidden predictive information from the large and/or complex databases. This technique uses past and present information to discover previously unknown patterns in the data, and then train and build models to predict future trends and behavior (Kantardzic, 2002). Classification trees are one of the most popular classification algorithms used in data mining. Classification trees are powerful knowledge models that predict the value of a target variable based on several input variables. They are easy to use, simple to understand and interpret, and require little data preparation (Tomei, 2008). Nevertheless, they do not always provide the most accurate result. A simple and effective procedure to tackle this deficiency is to use an ensemble of classifiers instead of using a single, large and less accurate tree classifier (Kuncheva, 2004). Classifier ensembles are aggregations of several classifiers (either different types of classifiers or different instants of the same classifier), whose individual predictions are combined in some manner (e.g., averaging or voting) to form a final prediction (Oza and Tumer, 2008). Because they use all the available classifier information, ensembles generally provide better and more robust solutions in most applications.

In this paper we investigate the application of such ensemble techniques to the classification and prediction of excess water production mechanisms in vertical oil wells. Excess water production is a serious economic and environmental problem in most mature oil fields. Accurate and timely diagnosis of the water production mechanism is critical in the success of the applied well treatment methodology. Incorrect, inadequate, or lack of proper diagnosis usually leads to ineffective water control treatments that cost a lot of time and money (Seright et al, 2001). Many empirical techniques such as decline curve plots, and water-oil ratio (WOR) versus cumulative oil production or time have traditionally been used in production data analysis (Poe

et al, 1999; Mohaghegh et al, 2005; Anderson et al, 2006). Typically, these plots have been used to determine whether the well is experiencing or likely to experience a water production problem; however, the significance of WOR in proper identification of the type of the water production problem in oil wells has not yet been fully investigated. Nevertheless, many oil companies to date apply the WOR diagnostic plots (Chan, 1995) for water production studies and problem diagnosis (Al Hasani et al, 2008; Sanchez et al, 2007). Although, in specific circumstances, these plots might help in diagnosing between two more common types of excess water production problems—namely coning and channelling—they are by no means general and applicable in all conditions (Seright, 1998). Without taking into account other important reservoir parameters, the WOR diagnostic plots could easily be misinterpreted. Studies conducted by our group on a series of simulated coning and channelling models also demonstrates deficiency of the WOR diagnostic method (Rabiei et al, 2009).

Considering the widespread use of water and oil production data in various reservoir investigations, it is plausible to assume that valuable information could be extracted from these data using modern mathematical and intelligent techniques. Production data are routinely collected and the accuracy of these data is usually reliable as most governmental regulatory agencies require accurate reporting of these values and also because these data represent revenue for oil companies.

In this paper we demonstrate the successful application of a sophisticated intelligent ensemble classifier, called random forest (Breiman, 2001), in classifying a number of excess water production mechanisms—namely channelling, coning and gravity segregation—based on WOR data. This classifier integrates various static reservoir parameters into production data (WOR) to reveal how the hidden predictive information in these data can be used to identify water production mechanisms with high accuracy rates. Unlike conventional WOR diagnostic studies that focus on the trends of log/log plots of WOR and the derivative of WOR against time, this study adopts a different approach to using WOR data.

Instead of plotting the WOR against time, we explore plots of WOR against the oil recovery factor (a dimensionless time, which is a ratio of cumulative oil being produced versus oil in-place with a maximum of unity) and extract predictive data points from these plots to be used in the ensemble classifier along with other reservoir parameters. Dimensionless groups are commonly used to generalise problems or plots (Seright, 1998). Yortsos et al (1999) applied a dimensionless time (fraction of pore volumes injected) for interpreting water/oil ratio in water floods. The use of dimensionless time will enable better analyses and comparisons of the WOR curves between models with a wide range of drainage areas and well operational histories, etc. The best estimate from a deterministic method or the P50 value from a probabilistic method can be used to estimate the oil recovery factor for the WOR plots. Alternatively, when the oil in-place estimate is not available, the conventional plots of WOR against time can be used to extract time-dependant predictive WOR points to be used in the ensemble classifiers; however, because of the above mentioned reasons, the use of oil recovery factor is preferred.

When reservoir parameters are not available, information from WOR data can be used alone. The frequency of high classification accuracy of more than 90% in our results demonstrates the significance of the use of production data in classifying the mechanisms of excess water production problem and proves that ensemble classifiers could be efficiently used to reveal the valid association between WOR data and the water production mechanism.

The rest of the paper is organised as follows: in the next section we briefly explain various reservoir models simulated for this study and describe how these models are used to generate a database to be used in the random forests algorithm. Next we give details on the predictor and target variables and the process of building pre- and post-production classifier models. Finally, we present the result and draw some conclusions.

METHODOLOGY

Classification algorithm

Ensemble classifiers are widely used in various fields and have been reported to improve the classification accuracy (Breiman, 1996, 2001; Bauer and Kohavi, 1999). Ensemble methods create a collection of prediction/classification models by applying the same algorithm on different samples generated from the original training sample, then make final predictions by aggregating (voting) over the ensembles (Pham, 2006). The random forests technique, developed by Brieman (2001), is a combination of tree-structured classifiers, in which each tree is grown in accordance with a random selection of input variables of a random sample drawn from the original training data by replacement. The output of the classifier is determined by a majority vote of the trees. The pseudo code for random forests is as follows:

- Suppose there are K cases in the training data; sample K cases at random by replacement from the original data;
- If there are M input variables, a number $m<<M$ is specified such that at each node of the tree, m variables are selected at random out of the M and the best split on these m variables is used to split the node;
- Each tree is grown to the largest extent possible; and,
- Make predictions according to majority vote of the set of K trees.

The random forests technique not only has a very good predictive performance but also has a number of other interesting features that make it worthwhile to investigate its applicability in various fields of science and engineering. It gives an estimate importance measure for each variable indicating the impact of each variable on the dependant variable. This technique also gives information about the relation between the variables and the classification and offers an experimental method for detecting variable interactions. Furthermore, random forests methodology is

tolerant of missing values, is easy to use and has a reasonable computing time (Lariviere and Poel, 2005).

Reservoir models

Random forests classifier requires a training dataset from which random trees are grown. To provide this training dataset, synthetic reservoir models were built to simulate excess water production due to coning, channelling and gravity segregated flows. Water coning occurs when the water/oil contact locally rises toward the completed interval of a well that normally produces from an oil column lying on top of an active aquifer (Seright, 1998). Water channelling is common when high permeability layers or fractures allow early water breakthrough during water flooding (Seright, 1998). These models were simulated using a commercial reservoir simulation software (Roxar, 2007).

BOTTOM WATER DRIVE (WATER CONING)

A radial model with a drainage area of 160 acres was built to simulate an oil reservoir with a water coning problem (Fig. 1). The radius of this sector model is 1,490 ft with a total thickness of 300 ft. The oil column is 100 ft thick with a 200 ft water column. This model is populated with a constant porosity of 0.2 and a constant permeability of 1,000 mD. A vertical well is perforated at the top 20% of the 100 ft oil column with a wellbore radius of 0.25 ft. To model a strong aquifer, the radial grids of the bottom five layers (100 ft in thickness) that are farthest from the vertical well have a porosity value of 2,000. The resulting aquifer is ~12 times the oil pore volume.

EDGE WATER DRIVE (WATER CONING)

A 3D Cartesian grid, with an area of 1,500 ft x 1,700 ft, a true vertical thickness of 100 ft and a dip angle of ~5°, was built to simulate water coning caused by an edge aquifer drive (Fig. 2). It is a homogeneous model with a constant porosity of 0.2 and a constant permeability of 1,000 mD. A vertical well is perforated to a true vertical depth of 50 ft above the oil-water contact (OWC).

WATER INJECTION (WATER CHANNELLING)

To simulate water channelling due to water flooding, three scenarios for small and large drainage area with different combinations of flow units are considered. The first scenario consists of a 3D Cartesian grid of 1,000 ft x 1,000 ft with a reservoir thickness of 100 ft (Fig. 3). An injector and producer pair is placed at both ends of this sector model for a direct line-drive water flooding pattern. To simulate models with little cross-flow between layers, the flow units are separated by low permeability layers of 10 mD and 5 mD. For models with cross-flow or high Kv/Kh ratios, these low permeability layers are removed, so a gravity dominated flow is observed between the injector and the producer. Despite a difference in permeability, this model has a constant porosity of 0.2 for all flow units.

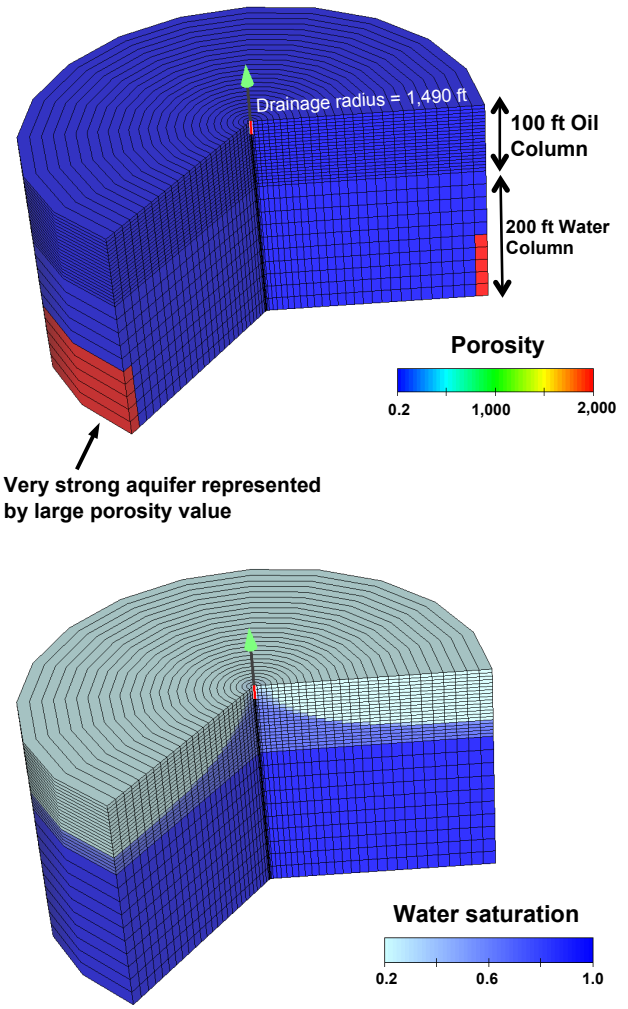

Figure 1. Symmetrical water coning by strong bottom aquifer.

The second scenario has the same settings as the first scenario except for the drainage area. A Cartesian grid with a larger drainage area of 1,500 ft x 1,700 ft was built to simulate a water injection scenario with a larger sweep area.

The same grid as in second scenario was used to simulate a third scenario with a different combination of the flow units. This model has four flow units of 3,000 mD, 1,000 mD, 500 mD and 2,500 mD from top to bottom layers. Similar to the first scenario and to represent little cross-flow between layers, the four flow units were separated by two magnitude lower permeability layers. A constant porosity of 0.2 is used, although the permeability is not constant.

EDGE WATER DRIVE (WATER CHANNELLING)

The 3D Cartesian grid is the same as the water coning (edge water drive) model, but the vertical well is placed farther up dip (Fig. 4). The grid dimension is also different, as the grid blocks are refined around the well and coarsened in the aquifer and farther away from the well. The strong down-dip aquifer provides the energy to sweep

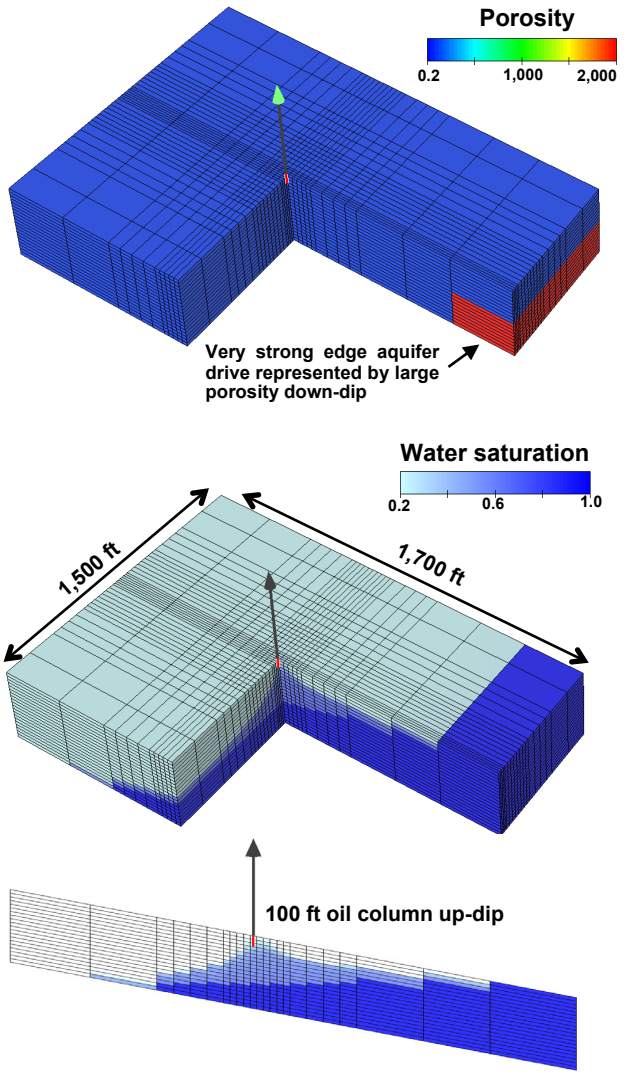

Figure 2. Asymmetrical water coning by edge aquifer drive from down-dip.

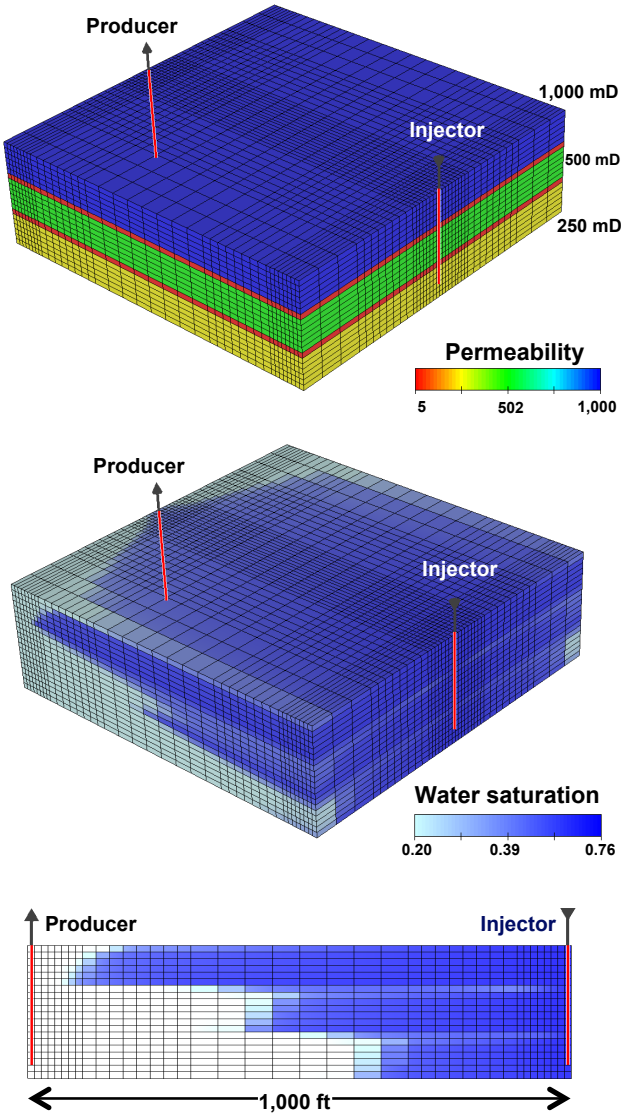

Figure 3. Water channelling caused by water injection.

the oil toward the up-dip producer. Three flow units of 6,000 mD, 1,500 mD and 500 mD from top to bottom are modelled as shown in Figure 4. The difference in permeability is set to allow the high permeability flow units to have a distinct water breakthrough. A constant porosity of 0.2 is used throughout this model.

Another scenario with four flow units of 6,000 mD, 500 mD, 2,000 mD and 200 mD from top to bottom is also simulated.

BOTTOM WATER DRIVE WITH BAFFLES IN VERTICAL DIRECTION

A 3D Cartesian model was built to simulate the water production from a reservoir with baffles in the vertical direction (Fig. 5). In this model spherical thin impermeable layers (800 ft in diameter) were randomly populated to act as zero transmissibility in the vertical direction. As shown in Figure 5, the thin impermeable zero vertical transmissibility spheres were modelled to provide baffles for the encroaching bottom water. It was observed that cone forming is minimal and this model exhibits a channelling behavior.

Dataset generation

A total of 716 cases of the aforementioned models were generated by varying a number of input parameters in each model. We observed that some of the water injection and edge water drive models with high Kv/Kh ratio and low permeability layers showed the gravity dominated flow or water under-run problem (Bailey et al, 2000). These cases are labelled GravityDominated in our study. Those cases where water production rate does not reach the defined critical point of WOR equal to 0.1 (or 9% water cut) in our analysis are labelled NoWater cases. These cases are used

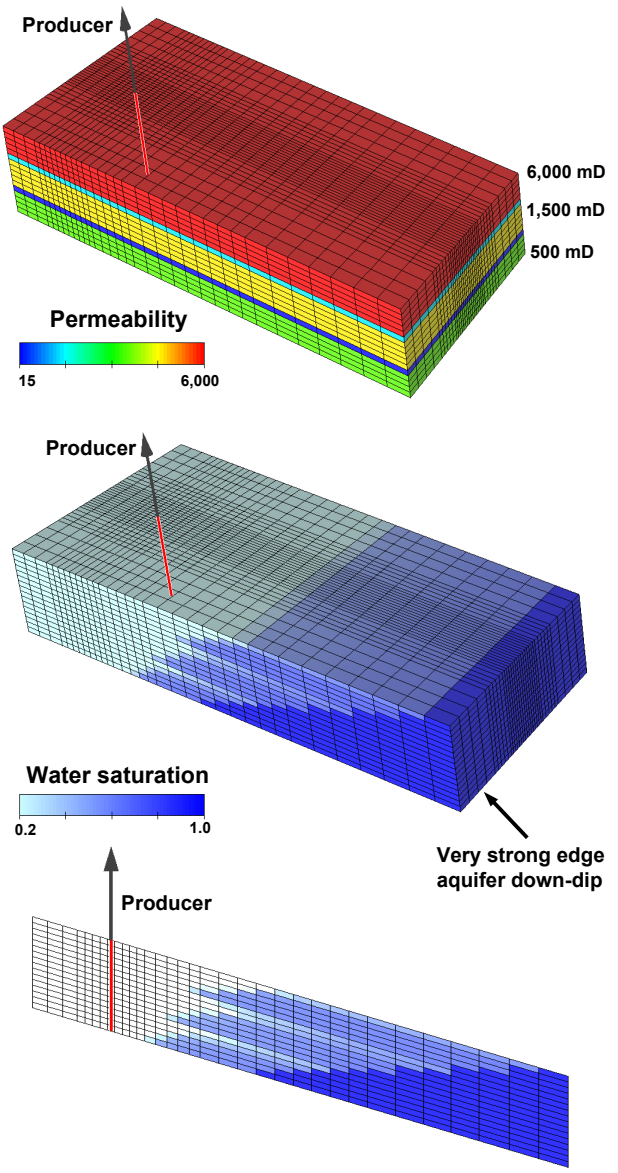

Figure 4. Water channelling caused by edge water drive.

as the control cases for investigating the efficiency of the classifier in identifying risk free situations.

A training subset containing roughly two thirds of the cases in this dataset was randomly chosen for building classifiers. The remaining cases were used for evaluating the performance of the trained classifiers.

Static and dynamic predictor variables

Each case in the database is defined by a set of static and dynamic variables. The static variables represent reservoir characteristics commonly available and are chosen from the input parameters used for generating various cases from the simulation models (Table 1).

We are also introducing new dynamic predictor variables to be used in water production problem analyses,

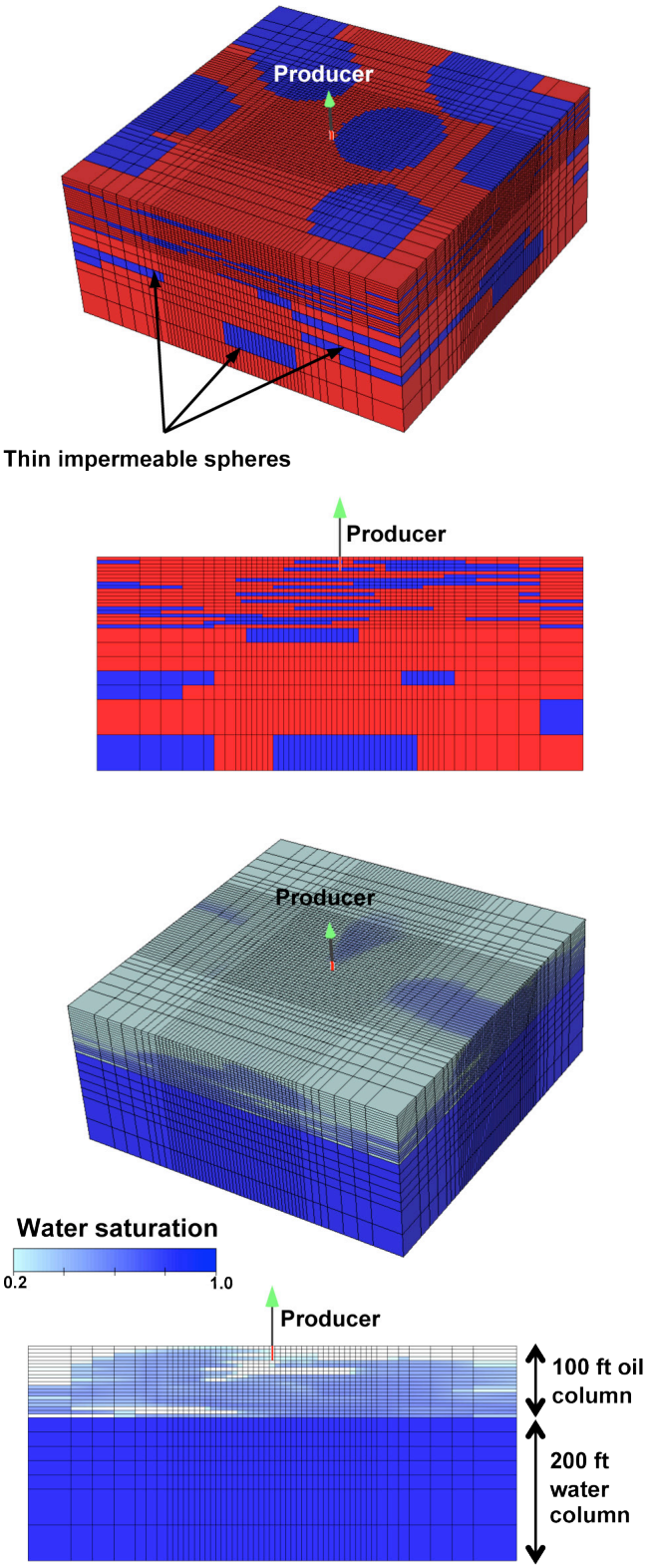

Figure 5. Water channelling through baffles in vertical direction from bottom aquifer.

which are obtained from the WOR versus the oil recovery factor (RF) plots. Using heuristic studies, some pilot points representing the characteristics of the WOR data are extracted by segmenting these plots at certain intervals where gradient remained constant (Fig. 6). These points are denoted as $RF_{WOR(0.1\ to\ 40)}$ (e.g. $RF_{WOR0.1}$ represents the value of RF at WOR equal to 0.1). The RF values below the point of WOR 0.1 are too small to yield any helpful information and hence are discarded. The cut off value point is at WOR equal to 40, which represents 97.5% water cut. A total of 14 RF_{WOR} variables were selected in this manner as dynamic variables. The analysis of variance (ANOVA) technique was used to assess the significance of the chosen variables with respect to identifying water production mechanism/problem type.

Random forests models

Classification algorithms were generated using a package in R software (R Development Core Team, 2009) called randomForests. The cases in the training subset were used to build and train the randomForests classifier. The first model is a control model in which only static variables are incorporated without including any dynamic production data. Such a model could be used before the start of production to get a rough estimation of the possibility of excess water production in the well. This model will be referred to as Model #0 throughout this paper.

- Model #0: static variables only

Additional models were also generated by integrating production data into Model #0. To be able to thoroughly examine the effect of the proposed dynamic variables, it was decided to implement a separate classifier for each dynamic predictor variable while taking into account the history of WOR trends before that specific point; i.e., 14 models were trained in the manner summarised below:

- Model #1: all static variables + $RF_{WOR0.1}$ + $RF_{WOR0.5}$
- Model #2: all static variables + $RF_{WOR0.1}$ + $RF_{WOR0.5}$ + RF_{WOR1}
- Model #3: all static variables + $RF_{WOR0.1}$ + $RF_{WOR0.5}$ + RF_{WOR1} + RF_{WOR2}

.
.
.

- Model #14: all static variables + $RF_{WOR0.1}$ + $RF_{WOR0.5}$ + RF_{WOR1} + RF_{WOR2} + … + RF_{WOR40}

Another set of models were also built using dynamic RF_{WOR} values alone without any static parameter.

- Model #1*: $RF_{WOR0.1}$ + $RF_{WOR0.5}$
- Model #2*: $RF_{WOR0.1}$ + $RF_{WOR0.5}$ + RF_{WOR1}
- Model #3*: $RF_{WOR0.1}$ + $RF_{WOR0.5}$ + RF_{WOR1} + RF_{WOR2}

.
.
.

- Model #14*: $RF_{WOR0.1}$ + $RF_{WOR0.5}$ + RF_{WOR1} + RF_{WOR2} + … + RF_{WOR40}

These models will be used to demonstrate whether production data (WOR) alone can reveal any helpful information in identifying water production mechanisms.

It is worth mentioning that the WOR ratios for some of the cases in our synthetic database do not reach to RF_{WOR40}. Depending on the strength of the water injector or aquifer, the maximum WOR values can be far less than 1 (50% water cut). In other words, it means that there are fewer cases with higher WOR points in our database.

Table 1. Input parameters for dataset generation.

Variable Name	Abbreviation
Vertical to horizontal permeability	Kv/Kh
API	API
Wettability	WET
Initial oil flow rate (normalised using the oil in-place)	IOFR
Plateau period for the initial oil flow rate	PP
Drainage area	DA
Aquifer strength–water/oil volume	AQWOV
Water injection rate	WIR

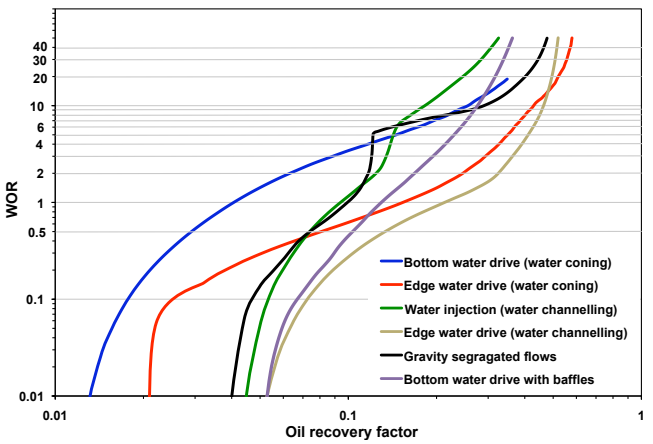

Figure 6. Sample plots of WOR against oil recovery factor for different simulated reservoir models.

VALIDATION DATA

The learnt patterns from the cases in the training dataset are applied to the remaining cases in the database to evaluate the efficiency of the trained classifier in classifying each case into one of Coning, Channelling, GravityDominated or NoWater categories. The performance of each implemented model is evaluated by applying it to the validation dataset and predicting the type of the water production problem for each case. The efficiency of each model is evaluated based on the percentage of correctly classified cases, which is calculated by dividing the total number of correctly classified cases by the number of cases tested in each model. We also use kappa coefficient (Cohen, 1960) to evaluate the efficiency of the models. Cohen's kappa measures the classifier's performance while taking into

account those successfully classified cases that might be attributed to chance alone.

RESULTS AND DISCUSSION

In this section, classification results of the implemented models are presented for three stages: the pre production stage which uses Model #0 and post-production stage which uses Models #(1–14) and Models #(1*–14*).

Pre-production stage

The model implemented with only static variables, was tested on the validation data to evaluate how effective reservoir characterisations alone can be used for classification purposes. As explained earlier, the random forests technique produces a combination of tree-structured classifiers, and a majority vote from these classifiers is the final output of the random forests classifier. One of these tree-structured classifiers used in the static model is shown in Figure 7 as an example. In Figure 7, each node represents a variable with a split value written below the variable name. At each node, cases that have variables with lower values than the split value go to the left of that node and the rest go to the right. For categorical variables, the split value is a level in each category; if a case has a categorical variable, which has the same value as the split value, it goes to the left, and otherwise it goes to the right side. This process continues until all cases are allocated to one of the terminal nodes highlighted in blue, which denote the predicted category for that case.

The overall classification accuracy rate for this model is 88% (see Table 2). Although, the classifier performs well in identifying channelling and coning cases individually with accuracy rates of 88% and 93% respectively, we see a less impressive result for risk free cases and an even lower accuracy rate for gravity segregated cases.

Post-production stage: production data plus reservoir parameters

The overall accuracy rates of the post-production models are presented in Figure 8. Detailed classification rates for each category are shown in Table 2. Our results demonstrate that once the production data is introduced to the models we start to see improvements in problem identifications. This improvement for the GravityDominated category is gradual, while we see instant enhancement in the accuracy rates for the Coning and NoWater categories with 6% and 16% increase respectively. It is observed that accuracy rates for the Channelling cases start to decrease at the earlier stages of water production; however, as more water production data becomes available, the classification rates start to increase and reach a level comparable to the static model and higher (91%). The relatively low percentage of correctly classified GravityDominated cases could be because of the similar behavior to the channelling problem and we anticipate that including more reservoir parameters might improve the result. As shown in Table 2, by including production data in the classifier, the coning problem can be identified with a staggering confidence level of more than 98%. Similarly, all cases with no excess water production are correctly identified, which means compared to the static model we observe 16% improvement in diagnosis.

Post-production stage: production data without reservoir rarameters

Another set of models developed in this study are classifiers generated using only production data without knowing any reservoir parameters. As shown in Table 2, although the overall accuracy rates of these models are not comparable to the models discussed earlier, with at least 85% accuracy rate, they can still provide a valuable tool for preliminary investigations for production engineers. They can be applied when an immediate assessment of the situation at hand is required and/or where none or little reservoir information is available.

Discussions

The results from various models implemented in this study corroborate the general idea that knowing the reservoir characterisation would help to some extent in determining whether the well is likely to experience water production or not, but when it comes to identifying specific problems such as gravity segregation, it is clear that reservoir characteristics alone are not very helpful. It has been shown that including the proposed dynamic production variables greatly enhances the diagnosis accuracy especially in identifying risk free situations.

The results summarised in Figure 8 represent the overall percentage of correctly classified cases in our validation subset. A more robust measure to show the efficiency of the proposed classification technique is the kappa coefficient. Kappa values higher than 0.8 are usually considered as very good agreement, disproving the role of chance. As can be seen in Figure 9, kappa coefficients for all the dynamic models with reservoir parameters are very high, which confirms the efficiency of the proposed classification algorithm. Dynamic models without reservoir parameters have slightly lower kappa coefficients but they are still in a range considered as good agreement. The static model also shows good agreement with a kappa of 0.8 (not shown in Figure 9).

The consistency of the proposed models in allocating each validating case to one of the classification categories is also explored by producing a bar-plot of the predicted class of each case with regard to the model used (Fig. 10). It is observed that the outputs of Models #(1–14) for each case are comparable and we see a consistency in the predicted class for most of the cases regardless of the model being used. It is clear that as more water production data is used in models, most of the misclassified cases are correctly classified.

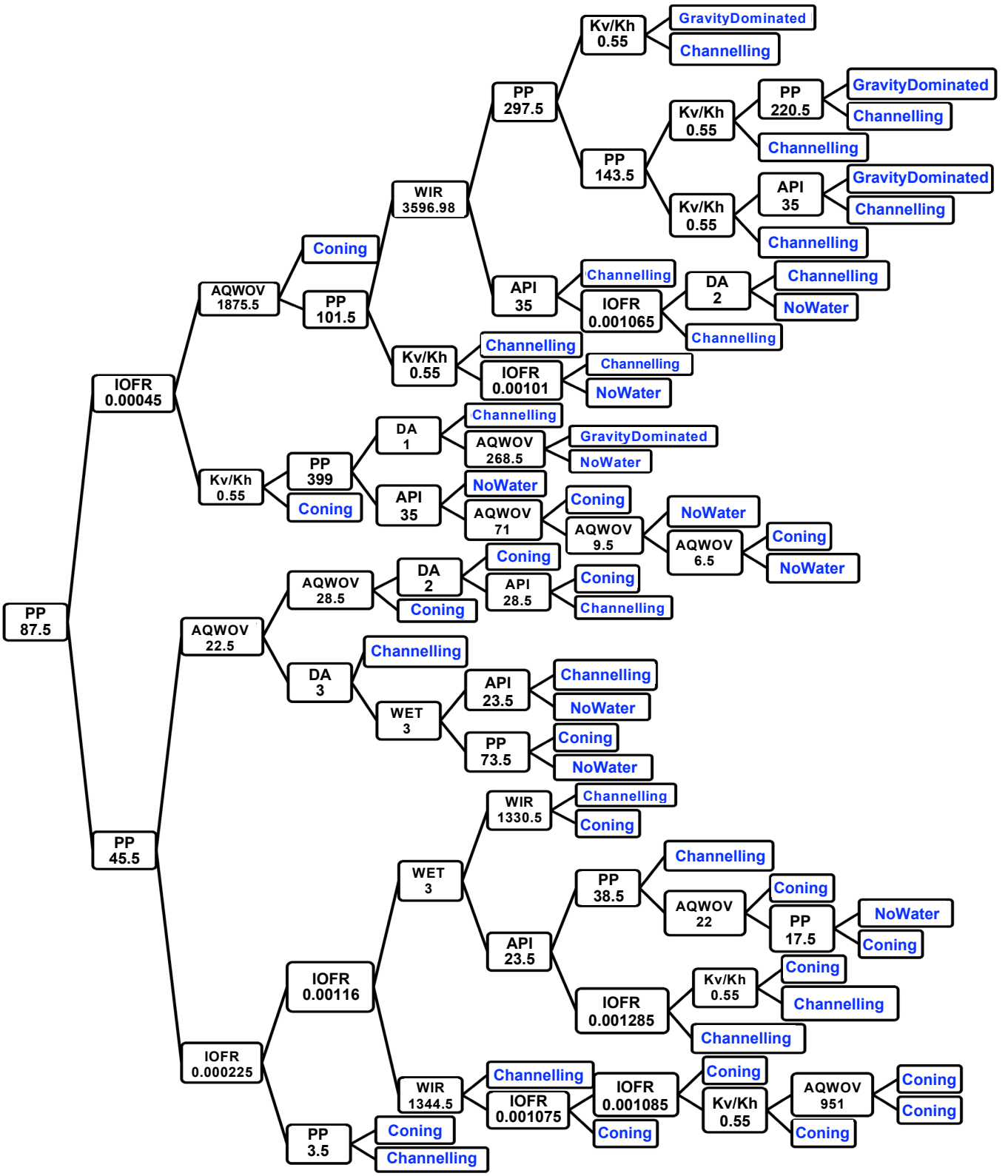

Figure 7. A schematic of one of the classifiers used in the random forests models.

Table 2. Individual and total accuracy rates for pre- and post-production classification models.

Model	Channelling	Coning	GravityDominated	NoWater	Total Accuracy (%)
Pre-production model	88%	93%	59%	84%	88%
Model #1	83%	99%	59%	100%	92%
Model #2	78%	99%	59%	100%	90%
Model #3	80%	98%	65%	100%	90%
Model #4	80%	100%	65%	100%	91%
Model #5	85%	98%	65%	100%	91%
Model #6	83%	98%	71%	100%	91%
Model #7	83%	98%	71%	100%	91%
Model #8	83%	98%	71%	100%	91%
Model #9	85%	100%	71%	100%	92%
Model #10	87%	98%	71%	100%	92%
Model #11	87%	98%	71%	100%	92%
Model #12	87%	98%	71%	100%	92%
Model #13	87%	98%	65%	100%	91%
Model #14	91%	100%	65%	100%	93%
Model #1*	81%	93%	18%	100%	85%
Model #2*	87%	96%	24%	100%	88%
Model #3*	85%	95%	24%	100%	86%
Model #4*	83%	95%	29%	100%	86%
Model #5*	80%	96%	29%	100%	85%
Model #6*	83%	96%	35%	100%	87%
Model #7*	85%	96%	35%	100%	87%
Model #8*	85%	98%	35%	100%	88%
Model #9*	85%	98%	35%	100%	88%
Model #10*	87%	96%	35%	100%	88%
Model #11*	87%	96%	35%	100%	88%
Model #12*	87%	96%	47%	100%	89%
Model #13*	84%	93%	47%	100%	87%
Model #14*	93%	97%	47%	100%	90%

Finally, an interesting feature of the random forest algorithm called the variable importance measure, which shows the impact of each predictor variable on the dependent variable, is explored and the results are shown in Figure 11. This Figure shows the importance measure for the pre-production model and two of the post-production models with and without static variables. Using this measure we find evidence that when dynamic production data is introduced in to the model, they play a more important role in identifying problem type, while reservoir characterisation data are not disregarded either. Another important finding is that as the well produces more water the role of aquifer strength becomes more distinctive. On the other hand, the distribution of the variable importance values reveals that parameters like wettability and API are less powerful in terms of predicting excess water production mechanisms, although they have significant effects on the amount of the produced water. When no static variable is used, it is

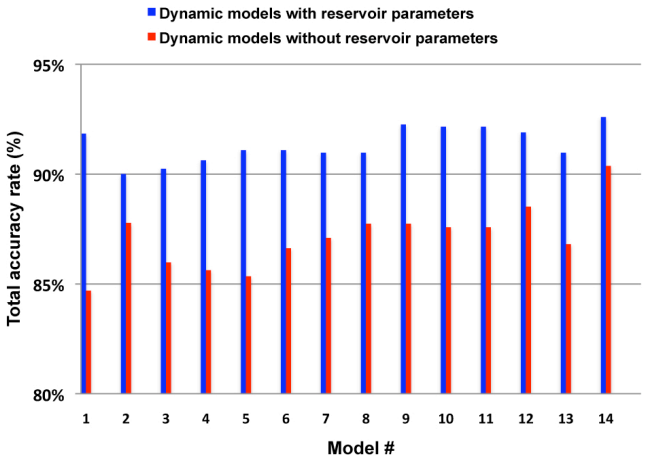

Figure 8. Total accuracy rates for dynamic classification models.

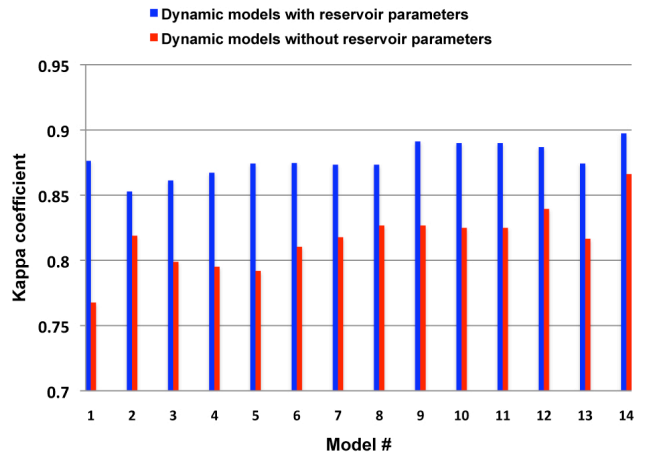

Figure 9. Kappa coefficient for dynamic classification models.

observed that the sequence of larger dynamic variable importance slightly changes, while smaller dynamic variables still have the dominant importance.

Our results demonstrate that dynamic models, which integrate static reservoir parameters and dynamic production data, outperform other models developed with only one of these data groups. While other models are useful for situations with limited data availability, we recommend using models with both static and dynamic variables to increase the precision of the diagnosis. The proposed models can be used at early stages of the problem forming with as little water cut as 30% to reasonably diagnose the type of the problem that is likely to happen. This means remedial actions could be taken before reaching a critical state later in the well's lifecycle.

The discussed models are not presented in this paper. For more information contact the lead author.

CONCLUSION

In this paper we demonstrated the significance of water and oil production data in diagnosing the mechanism of excess water production in vertical oil wells. A number of featured variables extracted from plots of WOR against oil recovery factor were introduced as predictor variables in a classification process. Synthetic reservoir models were used to generate a database from which a training subset was chosen to train random forests classification models. These models were then used to identify water production mechanisms in new synthetic cases in a validation subset

The findings reported here establish that the proposed classification technique could be successfully applied to predict water production mechanism based on reservoir characterisations with a reasonable accuracy rate. It is demonstrated that using the proposed featured WOR values would greatly enhance the performance of the classification technique. Our results reveal that WOR monitoring could also help in predicting the type of the water production mechanisms before the actual problem hitting the well, which means remedial actions could be taken accordingly ahead of time.

We have also shown that by using the proposed classification technique, WOR data alone without other reservoir characterisations, could also provide a valuable tool for preliminary investigations in oil fields.

We anticipate that the complete water production diagnostic system will be made available to the end user as a stand-alone tool in the future. The training data used in the present study will be updated as more real field data and different types of water production mechanisms are made available. Similarly, the classifier models will be redeveloped and updated using the new extended training dataset and will be validated through a range of techniques.

REFERENCES

AL HASANI, M.A., AL KHAYARI, S.R., AL MAAMARI, R.S. AND AL WADHAHI, M.A., 2008—Diagnosis of excessive water production in horizontal wells using WOR plots. SPE International Petroleum Technology Conference, Kuala Lumpur, Malaysia, 3–5 December, SPE 11958–MS.

ANDERSON, D.M., STOTTS, G.W.J., MATTAR, L., ILK, D. AND BLASINGAME, T.A., 2006—Production data analysis: challenges, pitfalls, diagnostics. SPE Annual Technical Conference and Exhibition, Texas, USA, 24–27 September, SPE 102048.

BAILEY, B., TYRIE, J., ELPHICK, J., KUCHUK, F., ROMANO, C. AND ROODHART, L., 2000—Water control. Schlumberger Oilfield Review, 12 (1), 30–51.

BAUER, E. AND KOHAVI, R., 1999—An empirical comparison of voting classification algorithms: bagging, boosting, variants. Machine Learning, 36, 105–39.

BREIMAN, L., 1996—Bagging predictors. Machine Learning, 24 (2), 123–40.

BREIMAN, L., 2001—Random forests. Machine Learning, 45, 5–32.

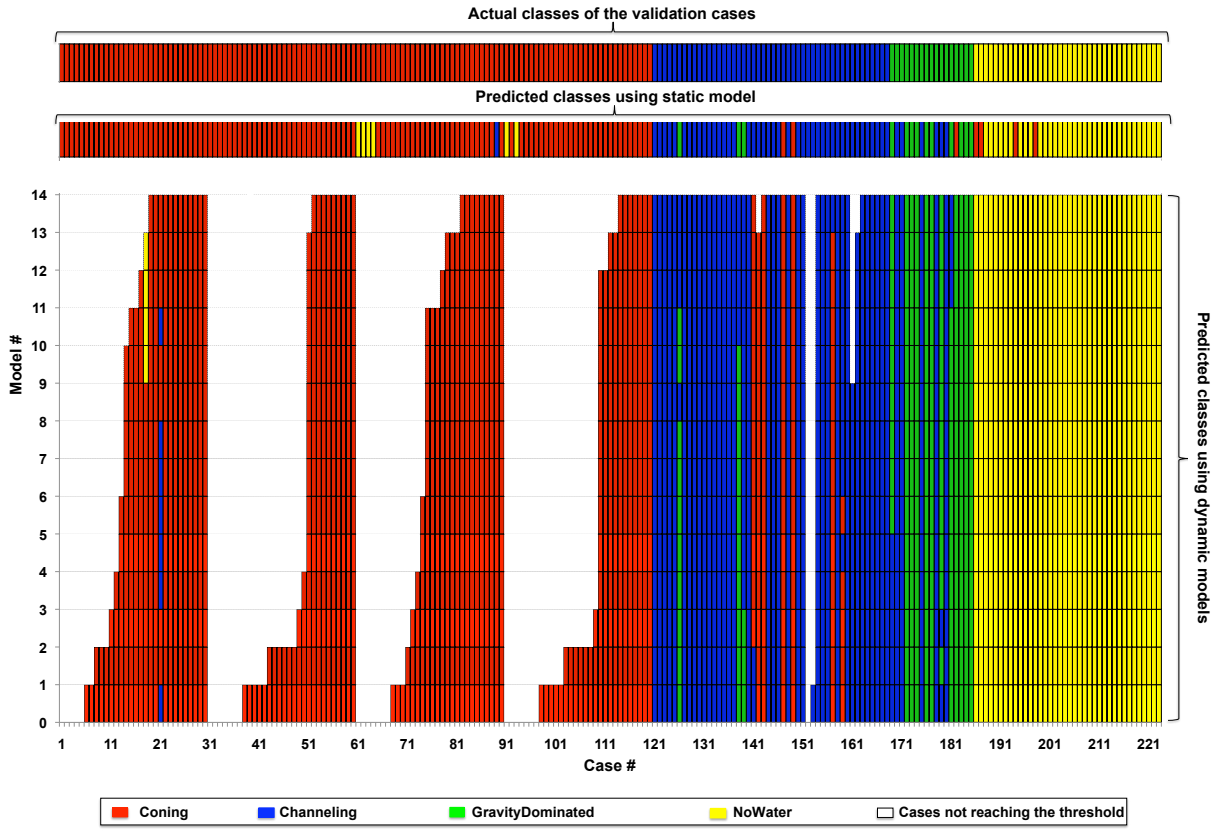

Figure 10. Sequential classification of each case predicted by each model compared to the known classes of cases.

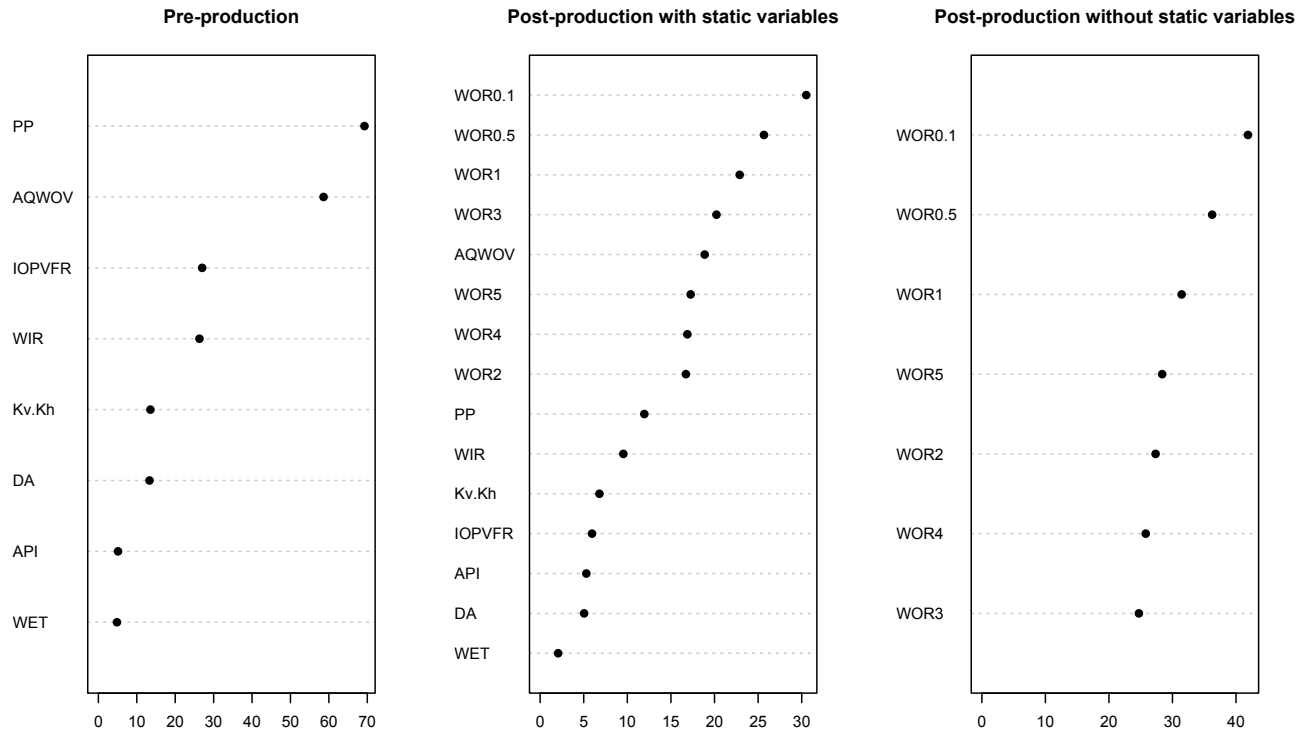

Figure 11. Variable importance plot for (a) pre-production, (b) past-production model with static variables, and (c) past-production model without static variables.

CHAN, K.S., 1995—Water control diagnosis plots. SPE Annual Technical Conference and Exhibition, Dallas, USA, 22–25 October, SPE 30775.

COHEN, J., 1960—A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 (1), 37–46.

KANTARDZIC, M., 2002—Data mining: concepts, models, methods and algorithms. USA: Wiley-IEEE Press.

KUNCHEVA, L.I., 2004—Combining pattern classifiers: methods and algorithms. USA: Wiley.

LARIVIERE, B. AND POEL, D.V.D., 2005—Predicting customer retention profitability by using random forests and regression forests technique. Expert Systems with Applications, 29 (2), 472–84.

MOHAGHEGH, S.D., GASKARI, R. AND JALALI, J., 2005—New method for production data analysis to identify new opportunities in mature fields: methodology and application. SPE Eastern Regional Meeting, Morgantown, USA, 14–16 September, SPE 98010.

OZA, N.C. AND TUMER, K, 2008—Classifier ensembles: select real-world applications. Information Fusion, 9, 4–20.

PHAM, H., 2006—Springer handbook of engineering statistics. Germany: Springer.

POE, B.D., ZHENG-POE, A. AND BONEY, C.L., 1999—Production data analysis and forecasting using a comprehensive analysis system. SPE Mid-Continent Operations Symposium, Oklahoma, USA, 28–31 March, SPE 52178.

RABIEI, M., GUPTA, R., CHEONG, Y.P. AND SANCHEZ SOTO, G.A., 2009—Excess water production diagnosis in oil fields using ensemble classifiers. International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 11–13 December.

SANCHEZ, P.Z.A., DELGADO, M.A. AND QUINONES, V.H., 2007—Water control in heavy oil mature field, Block 1AB. SPE Latin American and Caribbean Petroleum Engineering Conference, Buenos Aires, Argentina, 15–18 April, SPE 108039.

SERIGHT, R.S., LIANG, J.T., SCHRADER, R., HAGSTROM, J., LIU, J. AND WAVRIK, K., 1998—Improved methods for water shutoff. DOE/PC/91008-14, USA: National Technical Information Service, U.S. Department of Commerce.

SERIGHT, R.S., LANE, R.H. AND SYDANSK, R.D., 2001—A strategy for attacking excess water production. SPE Permian Basin Oil and Gas Recovery Conference, Midland, USA, 15–16 May, SPE 70067.

R DEVELOPMENT CORE TEAM, 2009—R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Accessed 29 September2009. <http://www.R-project.org>.

TOMEI, L.A., 2008—Encyclopedia of information technology curriculum integration. USA: Information Science Reference.

YORTSOS, Y.C., CHOI, Y., YANG, Z. AND SHAH, P.C., 1999—Analysis and interpretation of water/oil ratio in water floods. SPE Journal, 4 (4), SPE 59477.

THE AUTHORS

Minou Rabiei is a PhD student at Curtin University of Technology doing research in applied computing and statistics in excess water production problems in vertical oil wells. She completed a BSc at Tehran Amir Kabir University in 1998 and an MSc at London South Bank University in 2002. Her research interests are in the application of statistics and information technology in petroleum engineering. Member: SPE.

Ritu Gupta is a senior lecturer in the Department of Mathematics and Statistics at Curtin University of Technology. She obtained a Masters degree in 1989, followed by PhD in 1994, all in the area of statistics from Delhi University, India. She has worked in range of academic and consulting positions, first as lecturer in Delhi University, then as a research officer at the statistical consulting unit, University of Western Australia, and finally joined Curtin University of Technology in 2000. Ritu has written more than 30 consulting reports for various government departments and industries in Australia. She has presented at several international conferences, and has published papers in refereed journals. Over the past decade she has worked on developing statistical methods for efficient reserves estimation in petroleum engineering applications. This work has resulted in over 15 research papers, two PhD supervisions and software development. Accredited member: Statistical Society of Australia.

Yaw Peng Cheong is a reservoir engineer at CSIRO petroleum and geothermal research portfolio, carrying out reservoir simulation studies. Previously, he worked as a reservoir engineer at Sirikit oilfield, PTTEP Thailand and at Roxar, Malaysia, completing some consulting projects and technical support for static reservoir modelling and dynamic simulation software packages. Cheong has a PhD in petroleum engineering from Curtin University of Technology (2005), studying the experimental design and analysis methods in reservoir uncertainty studies. Before that, he completed a MSc in petroleum geoscience and a BSc in geology. Member: SPE.

Gerardo A. Sanchez Soto is a chemical engineer with a Masters degree (at ULA University, Venezuela) and PhD in chemical engineering (Birmingham University, UK). Gerardo worked in the production department of PDVSA-Intevep for 14 years. Gerardo has been involved in research areas such as: interfacial physical-chemistry associated to bitumen in water emulsions and two/three phases fluids, mixing technology and scale up processes, oil dehydration, drilling fluids and others aspects of drilling technology. Gerardo joined CSIRO in 2005 working on a software tool to automatically link reservoir and hydraulic commercial simulators. In 2007 he became subsea and well technologies stream leader supervising technically six research projects and working directly in hydrates in flow assurance, compact separation and relative permeability modifiers projects. He has produced 20 confidential reports, five papers and has been a co-author of seven patents in Intevep and one in CSIRO.

APPEA