Rock Permeability Forecasts Using Machine Learning and Monte Carlo Committee Machines

Simon Katz[1], Fred Aminzadeh[2], Wennan Long[2]*, George Chilingar[2] and Matin Lackpour[1]

[1]*Russian Academy of Natural Sciences, U.S. Branch, 101 S. Windsor Blvd., Los Angeles, California*
[2]*Petroleum Engineering Program, School of Engineering, University of Southern California*

Received November 17, 2016; Accepted December 18, 2016

Abstract: We developed new concepts of extended Monte Carlo cross validation and Monte Carlo committee machines. We subsequently used those concepts to predict permeability by linear regression and machine learning methods such as Neural Networks, Support Vector machines, and Regression Tree. Among the parameters we calculated using extended Monte Carlo cross validation are: root-mean squared error of individual forecasts, forecast bias, correlation between forecast and actual permeability, and forecast instability as a measure of sensitivity to perturbations of the training set. Output of Monte Carlo committee machines is constructed as the average of machine learning outputs generated from multiple versions of perturbed training sets. We observed that Monte Carlo committee machines produced high stability forecasts, while individual machine learning forecasts (e.g. a single ANN) were characterized by lower stability. Higher accuracy forecasts were achieved when we applied machine learning methods and linear regression using permeability models that included quantitative and categorical predictors and second-order interactions among the predictors.

Keywords: Permeability forecast, linear regression, machine learning, Monte Carlo cross validation, committee machines

1 Introduction

Current research in modeling and forecasting rock permeability includes a variety of methods and models. The structure of permeability models is defined by several factors which include a set of petrophysical parameters utilized as predictors, heterogeneity of lithology in the studied area, a model that defines relations between

**Corresponding author*: wlong@usc.edu

DOI: 10.7569/JSEE.2016.629519

permeability and predictors, and methods used for model construction. The means used for rock permeability forecasts are linear, log-linear regression, and machine learning methods such as neural networks and support vector machines. An excellent analysis of correlations amongst permeability, porosity, confining pressure, cementation, and grain size was done by AlHomadhi [1]. Results presented in this paper indicate the importance of factors other than porosity in reliable permeability forecast. Log-linear regression model for permeability with porosity, specific surface area, and irreducible fluid saturation were developed and analyzed by Chilingarian et al. [2] (See Addendum). The model was tested on data from several carbonate reservoir rock areas in the former USSR. One specific feature of the regression models utilized in that paper is the presence of second-order interactions. The inclusion of irreducible fluid saturation in the set of predictors led to a high correlation between actual and predicted permeability values. Generally, the advantage of linear regression permeability models is in their interpretability. Their weak point is their rigid structure. More flexible but more difficult to interpret are machine learning techniques, such as neural networks [3–6] and support vector machines [5, 7] used for permeability forecasts. Additional enhancement of the efficiency of machine learning permeability forecasts might be done with committee machines [4, 8–11]. Efficiency of forecast is estimated with different versions of cross validation. Cross validation version 'leave one out' was used by Gholami [7] to validate the permeability model built with support vector machines. Monte Carlo cross validation was proposed in Qing's paper [12] and used for estimating the number of components in the calibration model. General review of machine learning methods, such as neural networks and soft computing for reservoir characterization, was presented by Nikravesh and Aminzadeh [13].

Authors of this paper concentrated on development and testing of two algorithms: (a) extended version of Monte Carlo cross validation and (b) algorithm for Monte Carlo committee machines. Although we will not delve into the details of the ANN committee machines, we will include a brief description from Aminzadeh and de Groot [11] in Appendix 2.

Classic methods of cross validation, such as leave one out or k-fold have estimated single parameter - forecast error. The extended version of Monte Carlo cross validation is aimed at detailed analysis of the forecast and includes estimation of additional parameters such as forecast bias, forecast instability, correlation between forecast and target parameters, and comparative analysis of accuracy of individual and committee machine forecasts. Developed methods and algorithms have wide applications for forecasts of different types of reservoir characterization. They may be utilized, for example, for analysis and forecast of reservoir porosity distribution and in the forecast of trends in hydrocarbon production [14].

In this paper, the analyses of forecasts were obtained using a dataset [15] published as open source at the pubs.usgs.gov website. That dataset includes data from Jurassic sandstones collected from 15 wells in the Norwegian sector of the North Sea in the depth range of 3300–4250 m. Each record in the dataset contains a

DOI: 10.7569/JSEE.2016.629519

permeability value and respective values of quantitative and categorical parameters. The total number of records with no missing values is 99. Quantitative parameters in the dataset are porosity, grain size and burial depth. Categorical variables in this dataset qualitatively characterize content of the following mineralogical components: microquartz, macroquartz, clay, and carbonates. They have the following discreet values: 0 - not observed, 1 - very minor, 2 - minor, 3 - medium, 4 - much, 5 - very much.

Authors of this paper use R rminer package [5] to run machine learning methods and R function lm() for linear regression.

2 Monte Carlo Cross Validation and Monte Carlo Committee Machines

In Monte Carlo cross validation the test set is formed as a set of records that are randomly removed from the analyzed dataset. The remaining records form a training set. A key element of Monte Carlo cross validation is the formation of the multiple randomly formed pairs (training set-test set) and generation of forecasts for each pair. If the number of randomly formed pairs (training set-test set) is large enough, any record in the analyzed dataset will appear in multiple test sets. As a result, multiple forecasts will be produced for each record in the analyzed dataset. One of the results of Monte Carlo cross validation is the average, over multiple forecasts, for an individual record in the analyzed dataset. The average over multiple forecasts may be interpreted as the output of the committee machine.

Monte Carlo committee machines designed and tested in this paper are analogs for bagging committee machines [16]. The authors designed MC committee machines in two applications: (a) compact description of Monte Carlo cross validation results, (b) high stability forecast not sensitive to possible perturbations of the training set. MC committee machines rely on the use of multiple randomly formed training sets constructed by removing a fraction of randomly selected records from the studied dataset. A test set may be formed as the union of the records randomly removed from the training set, so that its records will include values of predicted parameters. Then output of the committee machines will be part of the results produced by Monte Carlo cross validation. The test set may also be a predefined dataset with records that do not include values of a predicted parameter. In this case, the goal of Monte Carlo committee machines will be to produce forecasts of the parameter to be predicted in a new geologic area.

Output of the Monte Carlo committee machine is in the form below:

$$Y(k) = \frac{1}{N} \sum_{m=1}^{N} F(k,m) \tag{1}$$

where k is the index of the record in the analyzed dataset, m is the index of a randomly formed training set that was utilized to produce a forecast $F(k, m)$. N is

DOI: 10.7569/JSEE.2016.629519

the total number of randomly generated training sets that are used to produce individual forecasts for a record with index *k*.

Mean absolute bias (*mabF*) and mean absolute error (*maErF*) of an individual forecast for the set of records S and Bias of individual forecasts (*bF*(*k*)), absolute bias averaged over a set of records S, *mabF*(*S*), and averaged absolute error of individual forecasts, *mabErF*(*S*), are defined as:

$$bF(k) = P(k) - Y(K); \quad mabF(S) = \frac{1}{N(S)} \sum_{k \subset S} \left| bF(K) \right|;$$
$$mabErF(S) = \frac{1}{N(S)} \sum_{k \subset S} \left| P(k) - F(k,m) \right| \tag{2}$$

Root mean squared error of the forecast by the committee machine for the group of records is shown as:

$$RMSE(S) = \left(\frac{1}{N(S)} \sum_{k \subset S}^{N(S)} \left(P(k) - Y(k) \right)^2 \right)^{0.5} \tag{3}$$

where N(S) is the number of records in the subset S.

To analyze bias of the forecast produced by the committee machine we construct multiple committee machines and define mean output over multiple committee machine forecasts calculated for the same record:

$$Y_R(k) = \frac{1}{R} \sum_{r=1}^{R} Y(k,r) \tag{4}$$

where r is the index of the committee machine forecasts, R is the total of their number.

Bias of the forecast by the committee machine for the individual value of the forecasted parameter and mean absolute bias within a subset of analyzed records are defined by Eq. 4 and 5:

$$bCM(k) = Y_R(k) - P(k); \; mabCM(S) = \frac{1}{N(S)} \sum_{k \subset S} \left| bCM \right| \tag{5}$$

Instability index for individual forecasts is defined as:

$$instIF(k) = \frac{1}{M-1} * \sum_{m=1}^{M} \left| Y(k) - F(k,m) \right| \tag{6}$$

DOI: 10.7569/JSEE.2016.629519

Calculation of instability index for the Monte Carlo committee machine forecast, *instCM*(*k*) involves multiple committee machine forecasts *Yr*(*k*) (Eq.1); $1 \le r \le R$ using multiple Monte Carlo cycles.

$$instICM(k) = \frac{1}{R} * \sum_{r=1}^{R} \left| Y_R(k) - Y(k,r) \right| \tag{7}$$

where R is total number of Monte Carlo cycles used to produce R forecasts for records with index *k*.

Values of the above parameters depend on the parameter named 'perturbation index' that should be defined prior to the start of a Monte Carlo cycle.

$$prInd = 1 - \frac{n.record(Train)}{n.records(dataSet)} \tag{8}$$

where *n.records* (*dataSet*) and *n.records* (*Train*) are respectively the number of records in the analyzed data set that includes all available records with known value of predicted parameter and in the randomly formed training sets, Train in Eq. 8 is a randomly formed train set. This parameter satisfies the following constraints: 1> *prInd*>= 0, and for stable, not perturbed training sets *prInd* = 0.

Major application of the MC committee machines forecast with a fixed test set with records that do not include predicted parameter, so that perturbations are done only to the train set. In that case the goal of the Monte Carlo committee machine is to decrease instability of the forecast of predicted parameters.

3 Performance of Extended MC Cross Validation and Construction MC Committee Machines

This section presents a sequence of steps that form a complex algorithm of the extended MC cross validation and construction of MC committee machines. Its goal is analysis of accuracy of the forecasts, and an increase of forecast stability.

1. Assign two parameters: (a) number of MC runs in the whole MC cycle and (b) perturbation index.
2. Build randomly formed pair of train-test sets with their sizes assigned in accordance with value of perturbation index.
3. Perform MC run that produces predictions of the forecasted parameter using a pair of train-test sets. Save results of the predictions as the data-set that includes actual values of the predicted parameter and respective predictions.
4. Perform full MC cycle by repeating multiple times steps 2 and 3. If a number of individual MC runs in the cycle is large enough, multiple predictions

DOI: 10.7569/JSEE.2016.629519

will be produced for each value of forecasted parameter in the analyzed dataset.

5. Calculate forecasts by the committee machines for the record with index k as the average of individual forecasts for that record. Analyze accuracy and instability of individual forecasts, and accuracy of the forecast by the committee machines in accordance with equations 1 to 7.
6. Run multiple MC cycles to calculate the instability of committee machine forecasts.

   To use MC committee machines for forecasts with a predefined test set with records that do not contain a predicted parameter, keep the test set unperturbed and do perturbations only to the training set by randomly removing from it a certain number of records.

4 Parameters of Distribution of the Number of Individual Forecasts in Monte Carlo Cross Validation

Further, for this paper, we did Monte Carlo cross validation of permeability forecasts with 1000 Monte Carlo runs within each MC cycle. Therefore, 1000 pairs (train-test) are created for analysis of the forecast efficiency. The total number of records in the studied data set is 99. At each Monte Carlo run, 9 records are randomly assigned to the test; remaining records go to the training set. So perturbation index for the results presented herein is 0.091. Since each pair of train-test sets are randomly constructed, the number of individual forecasts for each record in the studied dataset is a random number. Figure 1 shows a histogram of a number of individual forecasts in Monte Carlo cycle of 1000 runs. Horizontal line marks 74 forecasts. One can observe that the number of individual forecasts is rather stable and, in this example, does not go below 74.

Parameters of distribution of the random number of individual forecasts, calculated for four MC cross validation cycles, are shown in Table 1. Each cycle had

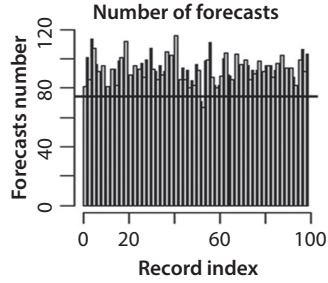

Figure 1 Histogram of a number of individual forecasts in Monte Carlo cycle of 1000 runs.

DOI: 10.7569/JSEE.2016.629519

Table 1 Parameters of distribution of a number of individual forecasts for individual records by MC cross validation in four MC cycles. Each MC cycle includes 1000 MC runs.

Parameters of distribution of the number of individual forecasts	Monte carlo cycles			
	1	2	3	4
Minimum	81	75	80	78
Lower quantile P: 0.25	94	95	96	93
Median	101	101	101	101
Mean	101	100	101	101
Upper quantile P: 0.75	108	107	107.5	108
Maximum	125	128	123	127

1000 cross validation runs. As seen, the estimated parameters of the distribution are similar across all four cycles. Minimum number of forecasts within four cycles varies from 75 to 81. Lower quantiles vary within a range of 93 to 96. Upper quantiles are in the range of 107–108. Therefore, if a number of Monte Carlo runs in the MC cycle is not less than 1000, one may expect that about 75 forecasts may be used to evaluate errors, bias and instability of the output of the committee machine and of individual forecasts.

5 Linear Regression Permeability Forecast with Empirical Permeability Models

The main advantages of multiple linear regressions when used for permeability forecasts are in fast data processing and in an explicit form of approximation permeability model. Conversely, the selection of appropriate permeability models and rigidity of the selected model may be problematic in the application of linear regression. The authors partially overcame these complications by considering several permeability models with different levels of complexity and different sets of permeability predictors. Complex models that include interactions among many predictors may over fit predicted variables. The tool for control of over fitting used in this paper is Monte Carlo cross validation. To be able to perform comparative analysis of the accuracy of linear regression and forecasts with several machine learning methods, we work with the same models for both linear regression and machine learning methods. These models are:

$$\text{Model 1: Permeability} \sim \text{Porosity} + \text{Grain.Size} \quad (9)$$

$$\text{Model 2: Permeability} \sim \text{Porosity} + \text{Grain.Size} + \text{Porosity*Grain.Size} \quad (10)$$

DOI: 10.7569/JSEE.2016.629519

Model3: Permeability ~ Porosity + Grain.Size + Depth + Microquartz + Porosity*Grain.Size + Porosity^Depth + Grain.Size*Depth (11)

Model 4: Permeability ~ Porosity + Grain.Size + Depth + Clay (12)

Model 5: Permeability ~ Porosity + Grain.Size + Depth + Microquartz (13)

Model 6: Permeability ~ Porosity + Grain.Size + Depth + Microquartz + Clay (14)

Model 7: Permeability ~ Porosity + Grain.Size + Grain.Size*Depth (15)

Model8: Permeability ~ Porosity + Grain.Size + Depth + Porosity*Grain.Size + Porosity^Depth + Grain.Size*Depth (16)

Model 9: Permeability ~ Porosity + Grain.Size + Depth + Microquartz + Clay + Porosity*Grain.Size + Porosity^Depth + Grain.Size*Depth + Porosity*(Clay + Microquartz) + Microquartz* (Grain.size + Porosity) (17)

Model 1 includes only two quantitative predictors with no interactions. Models of this type are often used in permeability forecasts with machine learning methods. Model 2 is an extension of Model 1 with interaction between two predictors in the models. Models 1, 2, and 7 include only quantitative predictors. Other models, in addition to quantitative predictors, include one or more categorical predictors that take discrete sets of values. They also interacted among predictors.

Individual forecasts produced by linear regression and machine methods may be negative, although predicted parameter in Eq. 9–17 is not smaller than zero. To improve accuracy of the forecast, all forecasts with negative values are assigned value 0.

Bias, instability, errors of individual linear regression forecasts and correlations of the output of MC committee machines with permeability are shown in Table 2. These parameters were estimated using MC cross validation with 1000 Monte Carlo cross validation runs in each MC cycle.

Values of mean squared error of forecast with all 9 models are about 80% larger compared to mean absolute errors. This might be due to the effect of outliers. According to Table 2, bias, individual forecast errors and instability are sizable for the simplistic Model 1. Values of these parameters for Models 3, 8 and 9 are significantly smaller than values of the same parameters obtained with Model 1. Thus, the effects of second-order interaction between porosity and Grain.Size and the effects of categorical predictors are significant.

DOI: 10.7569/JSEE.2016.629519

Table 2 Errors, bias, instability, and correlations of forecasts by MC committee machine with forecasted permeability.

Permeability models	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Mean Squared Error *RMSE* (*S*)	475	284	274	414	413	422	374	267	264
Mean Absolute Error *maErF*(*S*)	234	165	154	216	220	219	204	149	143
Mean absolute bias *mabF*(*S*)	228	165	147	215	221	212	201	144	142
Instability *instIF(k)*	37	21	28	37	38	39	33	27	31

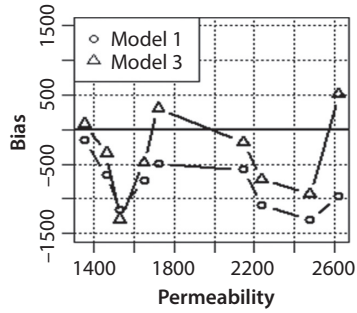

Figure 2 Bias of the forecast with permeability Models 1 and 3. Bias is calculated for permeability values in the range of 1100 -2650 mD. Horizontal scale – permeability.

Table 2 presents only partial information on the forecast accuracy produced by extended Monte Carlo cross validation. Important characteristics of the forecast accuracy are trends of accuracy characterization parameters as functions of a predicted parameter. This is illustrated by Figure 2, which shows bias of the individual permeability forecast as a function of the forecasted permeability.

Forecasts relying on the use of Model 1 have a large negative bias at bigger values of permeability, which may be as sizable as -1300 mD. Therefore Model 1 is not appropriate for forecast and identification of outsized permeability zones. Absolute values of the bias of the forecasts with Model 3 are smaller and not systematically negative. The difference in the properties of the forecasts produced with these two models is illustrated in Figure 3 which shows actual permeability values and their forecasts with two models. One can observe different signs of the bias produced with Model 1 and significant deviations of forecast values from actual permeability. As for forecasts with Model 3, they are much closer to forecasted permeability especially at large permeability values.

DOI: 10.7569/JSEE.2016.629519

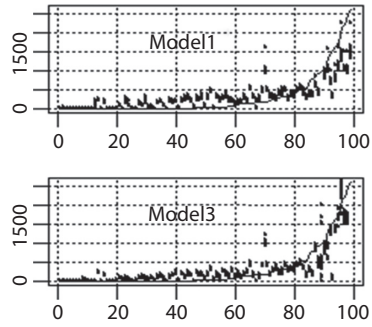

Figure 3 Ordered values of permeability and their forecasts with permeability Models 1 and 3. Continuous line - ordered permeability, small dots - individual forecasts. Vertical axis - permeability in mD horizontal scale - index of the record.

According to Figure 3, bias of the forecast with Model 1 is large and positive for permeability values smaller than 500 mD. It becomes negative for larger permeability values, so that permeability of 2600 mD is predicted as 1500 mD. A forecast with Model 3 is much more accurate and has smaller bias in a full permeability range.

6 Accuracy of the Forecasts with Machine Learning Methods

This section presents results of analyses of individual forecasts with six machine learning methods. These methods are: neural networks (NN), support vector machines (SVM), regression tree (RT), principal component regression (PCR), k-nearest neighbor (KNN), and regression with partial least squares (PLSR). Tables 3 and 4 show forecast bias and forecast instability for the same permeability models that were used for analysis of accuracy of linear regression. Both tables show similar patterns as those for linear regression forecast. The smallest bias was obtained for forecasts with Models 3, 8, and 9. Shown in Table 3, regression tree committee machines produced forecasts with the least bias when compared to other methods. Instability of forecasts by this method is not small and exceeds instability of linear regression forecasts. Instability of the individual forecasts is the largest for neural network. Hence, more individual forecasts may be necessary to build neural network or regression tree committee machine forecasts with low instability.

Smaller regression tree bias compared to the bias of support vector machines is illustrated by Figure 4. As seen here, bias of forecast by the committee machines relying on support vector machines is systematically negative at permeability values larger than 1500 mD. Therefore, a support vector machine is not suitable for forecasts of large permeability values. Bias of forecasts with regression tree is smaller and not systematically negative.

DOI: 10.7569/JSEE.2016.629519

Table 3 Mean absolute bias of the forecasts by Monte Carlo committee machines with six machine learning methods. Mean absolute bias is calculated for forecasts with nine permeability models for each forecast method.

	Methods					
Models	**PLSR**	**KNN**	**SVM**	**NN**	**PCR**	**RT**
1	231.9	153.0	169.5	152.1	230.3	118.8
2	164.0	154.5	166.5	153.7	164.0	121.1
3	146.0	139.7	143.6	126.5	146.7	124.7
4	217.6	142.7	142.2	127.0	215.9	125.1
5	221.7	157.3	146.3	133.3	221.7	121.9
6	214.8	147.7	155.7	143.5	215.0	123.5
7	205.4	151.6	164.4	137.0	204.3	120.5
8	144.2	133.4	141.4	133.4	143.2	125.4
9	143.1	145.5	142.7	148.5	142.2	123.3

Table 4 Instability index of individual forecasts with six machine learning methods.

	Methods					
Models	**PLSR**	**KNN**	**SVM**	**NN**	**PCR**	**RT**
1	30.6	47.8	89.6	127.7	31.6	95.7
2	14.8	45.3	84.9	131.6	14.9	107.8
3	21.4	45.1	82.2	173.8	22.2	105.8
4	31.7	48.4	88.8	163.1	33.0	102.2
5	36.6	51.6	81.9	142.0	32.3	95.7
6	34.0	49.9	93.4	168.2	30.7	94.1
7	25.6	48.1	81.9	141.9	28.0	93.4
8	21.7	40.4	83.0	156.6	23.2	98.2
9	28.6	49.6	84.8	209.1	30.2	94.6

DOI: 10.7569/JSEE.2016.629519

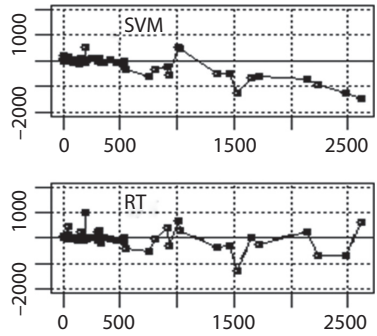

Figure 4 Forecast bias as a function of permeability values for two machine learning methods with permeability model (Model 1).

Table 5 Correlation coefficients between outputs of Monte Carlo committee machines and forecasted permeability.

	Machine learning methods					
Models	**PLSR**	**KNN**	**SVM**	**NN**	**PCR**	**RT**
1	0.65	0.86	0.86	0.88	0.67	0.910
2	0.88	0.86	0.86	0.89	0.87	0.911
3	0.89	0.85	0.90	0.91	0.89	0.909
4	0.72	0.85	0.90	0.89	0.72	0.909
5	0.73	0.83	0.90	0.90	0.72	0.911
6	0.73	0.85	0.90	0.89	0.73	0.910
7	0.79	0.84	0.86	0.89	0.78	0.909
8	0.89	0.87	0.90	0.90	0.89	0.913
9	0.89	0.84	0.90	0.90	0.89	0.913

Correlation between outputs of MC committee machines and forecasted permeability is shown in Table 5. The highest correlation coefficient at the level of 0.91 is for forecasts with regression tree. Regression adapts well to different permeability models so that even forecasts with simplistic Model 1 are characterized by a correlation of 0.91.

DOI: 10.7569/JSEE.2016.629519

7 Analysis of Instability of the Forecast

This section and the one following illustrate the importance of analysis of instability in individual forecasts. They also elucidate an increase of stability of the forecast by the MC committee machines via an increase of a number of individual forecasts that shape those committee machines.

Table 6 and Figure 5 present results of detailed analyses of accuracy and instability of NN forecasts for eight values of permeability. Individual forecasts were produced with permeability Model 2. The first row of the Table below shows actual values of predicted permeability, while other rows present results of analyses of

Table 6 Instability of individual forecasts for eight permeability records.

Parameters	**Output of MC cross validation**							
Index of the Record	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Permeability	755	808	917	933	1013	1019	1353	1461
MC committee machine forecast	486	712	493	436	1600	1886	1857	1078
Individual Forecast Error	272	398	581	606	1137	924	567	521
Maximum of Individual Forecasts	875	4624	2479	5655	6870	6743	4911	2573
Minimum of Individual Forecast	179	179	0	0	0	126	607	0

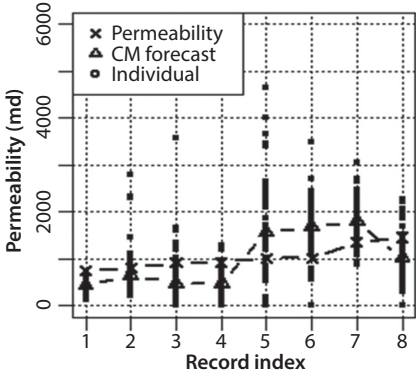

Figure 5 Individual and committee machine permeability forecasts in comparison with actual permeability.

DOI: 10.7569/JSEE.2016.629519

individual and committee machine forecasts. Committee machines output is built as an average of 75 individual neural network forecasts. According to Table 6, the range of individual forecasts is huge, and the difference between maximum and minimum individual forecasts for the same permeability value may be as large as 6000 mD. The committee machine forecast is closer to actual permeability.

Graphic illustration of instability of individual forecasts is given at Figure 5. Small dots denote the values of individual forecasts, which may even equal zero for actual permeability values exceeding 900 mD.

8 Enhancement of Stability of the MC Committee Machines Forecast Via Increase of the Number of Individual Forecasts

The effect of the number of forecasts that form the output of the committee machines on instability of the forecast is illustrated by Figure 6. It shows instability index of individual forecasts by support vector machines, *instIF* (*k*) in comparison to instability index, *instCM* (*k*) of the forecast by the SVM MC committee machines. The committee machine output is built as the average over several individual SVM forecasts. Values of *instCM* (*k*) are calculated for the records with permeability within the range of permeability greater than 500 mD. Parameter N at Figure 6 is the number of individual forecasts that form the SVM committee machine. The total number of committee machines R is used for the calculation of average $Y_R^{(k)}$. Results shown in Figure 6 were obtained for a small perturbation index equal to 0.1. According to this Figure, individual forecasts are sensitive to minor perturbations of the training set, with instability index of individual forecasts as high as 600 mD. The instability index of the committee machines declines with an increasing number of individual predictions in a committee machine. At 50 or more individual forecasts that form the committee machine, an instability index of the committee machines forecast is about 6 times less than that of individual predictions.

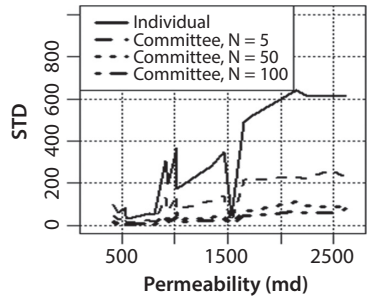

Figure 6 Instability indexes of individual permeability forecasts by support vector machines and by MC committee machines. N is the number of individual forecasts in a committee machine. Individual in Figure 5 is instability index of individual forecasts.

DOI: 10.7569/JSEE.2016.629519

9 Conclusions

New methods of extended Monte Carlo cross validation and Monte Carlo committee machines are introduced in this paper and tested with rock permeability forecasts. Extended Monte Carlo cross validation is designed for detailed analysis of bias and instability of the forecasts, forecast error, and correlation between forecast and forecasted parameter. Monte Carlo committee machines are expected to enhance forecast stability, which is improved with an increase of a number of individual forecasts that form the committee machine. Forecast bias is a basic characteristic of forecast accuracy. It cannot be diminished.

Comparative analysis of the accuracy of individual permeability forecasts and forecasts by Monte Carlo committee machines relying on linear regression, and several machine learning methods, was performed using extended Monte Carlo cross validation protocol. A list of analyzed machine learning methods includes neural networks, k-nearest neighbor, support vector machine, principal component regression, partial least squared, and regression tree. Committee machines relying on neural network and regression tree outperformed linear regression and are characterized by smaller forecast bias. The highest accuracy, smallest bias and highest correlation between forecast and forecasted permeability was obtained with regression tree.

Nomenclature

- MC cross validation = Monte Carlo cross validation, MC committee machine - Monte Carlo committee machine, MC run - Monte Carlo run.
- $F(k, m)$ = Individual forecast for record with index k with model built using randomly formed train set with index m.
- $mabF(S)$ = Mean absolute bias of individual forecast.
- $bCM(k)$ = Mean absolute bias of the forecast by the committee machine for the value of the forecasted parameter in the record with index k.
- $instIF(k)$ = Instability index for individual forecasts.
- $instCM(k)$ = Instability index for the MC committee machine forecast.
- $prInd$ = Perturbation index defining degree of random perturbations of the train set.
- SVM = Support vector machine, NN - neural net, KNN - k-nearest neighbor regression, PCR- principal component regression, PLSR - partial least squares regression, LR - linear regression, RT - regression tree.
- Notations for the models = Linear models for response variable for forecasted parameter are of the form:

$$responce = a_0 + a_1 * predictor_n + \ldots + a_n * predictor_n$$

DOI: 10.7569/JSEE.2016.629519

- Neither linear regression nor machine learning forecast functions in R accept models presented in this form. Instead, they use compact form of linear model.

$$responce \sim predictor_1 + ... + predictor_n$$

This form of linear model is utilized in this paper.
- MC run = Forecast for records in a single pair randomly generated train and test sets.
- Individual forecast = Forecast for individual record in the test set produced as part of Monte Carlo run.
- MC cycle = Full set of MC runs performed for analysis of accuracy of the forecast or construction of MC committee machine.
- Instability index = Forecast characterization parameter that defines potential changes of the forecast due to perturbations of the training set, such as a small increase of the size of the training set.

Appendix 1- Description of Permeability Models from Different Fields

Many carbonate reservoirs have very high irreducible water saturation. As much as 50–60% of pore space is occupied by this immovable water, which, for the most part, does not participate in the flow of fluids.

As the fluids move within the reservoir, they consider the water in the dead-end pores and fractures as being part of the solid rock. Thus, the definition of effective porosity that is used in many countries is incorrect. In the United States the definition of effective porosity is their "effective" porosity minus the irreducible fluid saturation. Russia defines it in the same way, but refers to it as open (inter-communicating) porosity. When we use the word 'effective' properly we will achieve more accurate results when assessing the correlation between porosity and permeability.

In addition, specific surface area (per unit of pore volume), which is a measure of the degree of fracturing, must be considered when evaluating the relationship between porosity and permeability. Fractures do not contribute much to porosity but they do substantially increase the permeability. A few near-perfect correlations were obtained by Chilingarian, Bagrintseva and Chang [2] for several carbonate reservoirs by adding two additional variables: irreducible fluid saturation (S_{wr}) and specific surface area (S_s):

1. Vuktylskiy Gas-Condensate Field, Russia

$$\log k = 0.9532 - 2.7880 * 10^{-2} S_{wr} - 5.5597 * 10^{-4} S_S + 1.3309 * 10^{-1} \varphi + 1.7707 * 10^{-5} S_{wr} S_S$$
$$(R = 0.997)$$

DOI: 10.7569/JSEE.2016.629519

2. Kuybyshev, Along-Volga Region, Russia

$$\log k = 2.1085 - 5.0777 * 10^{-2} S_{wr} - 4.3785 * 10^{-4} S_S + 7.9959 * 10^{-2} \varphi + 7.6326 * 10^{-6} S_{wr} S_S \quad (R = 0.998)$$

3. Orenburg Field, Russia

$$\log k = 3.4351 - 2.0442 * 10^{-1} S_{wr} - 9.5086 * 10^{-6} S_{wr} S_S + 8.0217 * 10^{-3} S_{wr} \varphi + 2.3892 * 10^{-5} S_S \varphi \quad (R = 0.981)$$

Where S_S = Specific Surface Area (per unit of pore volume), S_{wr} = irreducible water saturation and ϕ = fractional porosity. The role of the insoluble residue (IR) content is also being investigated by the writer in determining the relationship between porosity and permeability.

Appendix 2- A Brief Overview of Modular Networks or Committee Machines*

Modular Neural Networks (Committee Machines)*

Modular neural networks or dynamic committee machines are comprised of a cluster of individual neural networks, referred to as local experts, which are connected to each other. Usually they are controlled or governed by a global expert manifested by a gating network that acts as a referee or committee chair. The role of the global expert is to make rulings or judgments pertaining to the local experts, assign a significance factor or weight to their respective outputs, and to determine what role each should play in the final outcome of the combined network. The number of outputs of the gating network is the same as the number of individual networks that are combined. Figure 1–8 is an example of a network.

Dynamic committee machines are capable of handling more complicated problems. Examples of applications of committee machines in the oil industry include those in the inversion of induction of log data (Zhang and Poulton, [17] and prediction of shear wave logs from sonic and other suites of logs Fruhwirth and Steinlechner, [18].

Fruhwirth and Steinlechner [18] used a Completely Connected Perceptron (CCP) with active input neurons, a common MLP as well as an MLP with short cuts for log prediction application. They used the network to predict shear wave logs from a suite of other logs in a well. The training was done using an existing shear wave and other suite of logs from another well. The training started using

DOI: 10.7569/JSEE.2016.629519

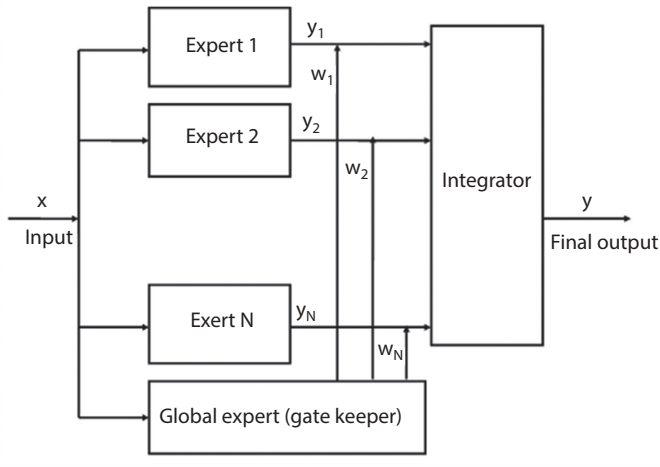

Figure A2- 1 Configuration of a Committee Machine.

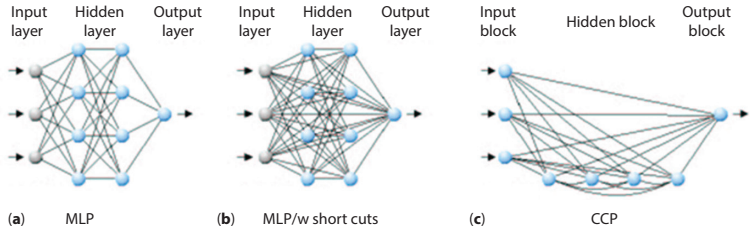

Figure A2-2 Multilayer Perceptron (**a**) without and (**b**) with short cut connections and passive input neurons (**c**) Completely Connected Perceptron (CCP) with active input neurons (from Fruhwirth and Steinlechner, 2005)

solely the CCP architecture for 10 network generations starting without any hidden units (Figure 1–9). This operation is equivalent to a multi-linear regression in many attribute analysis applications except that the nodes in CCP can account for any non-linear relationship between the known (in this case the known log suites) versus the unknown (the shear wave log).

The final CCP had 9 hidden units. In each of these generations 20 different and randomly initialized networks in parallel were trained, which can be considered another type of realization of a modular neural network. The goal was to prevent getting trapped too much in local error minima.

*This Appendix is adopted from Aminzadeh and de Groot [11]

DOI: 10.7569/JSEE.2016.629519

References

1. E.S. AlHomadhi, New correlations of permeability and porosity versus confining pressure, cementation, and grain size and new quantitatively correlation relates permeability to porosity. *Arabian Journal of Geosciences* **7**(7), 2871–2879 (2014).
2. G. Chilingarian, J. Chang, and K. Bagrintseva, Empirical expression of permeability in terms of porosity, specific surface area, and residual water saturation of carbonate rocks. *J. Petrol. Sci. Eng.* **4**(4), 317–322 (2006).
3. C.C. Fung, K.W. Wong, and H. Eren, Determination of a generalised BPNN using SOM data-splitting and early stopping validation approach, *Proceedings of Eighth Australian Conference on Neural Network (ACNN'97)*, pp. 129–133 (1997).
4. B. Alpann and H.B. Helle, Committee neural networks for porosity and permeability prediction from well logs. Geophysical Prospecting. **50**(6), 547–674 (2002).
5. P. Cortez and P. Perner, Data mining with Neural Networks and Support Vector Machines using the R/rminer Tool. Chapter in *Advances in Data Mining Applications and Theoretical Aspects*, Volume 6171 of the series Lecture Notes in Computer Science, 572–583 (2010).
6. W. Long, D. Chai, and F. Aminzadeh, *Pseudo Density Log Generation using Artificial Neural Network*. Society of Petroleum Engineers (2016, May 23). doi: 10.2118/180439-MS.
7. R. Gholami, A. Shahraki, and A.M. Paghaleh, Support vector regression for prediction of gas reservoirs permeability. *J. Min. Environ.* **2**(1), 41–52 (2011).
8. C-H. Chen and Z-S. Lin, A committee machine with empirical formulas for permeability prediction. *Computers & Geosciences* **32**(4), 485–496, ISSN 0098-3004 (May 2006).
9. R. Sadeghi, A. Kadkhodaie, B. Rafiei, and S. Khodabakhsh, A committee machine approach for predicting permeability from well log data: A case study from a heterogeneous carbonate reservoir, Balal oil Field, Persian Gulf. *Geopersia* **1**(2), 1–10 (2010).
10. L. Heng and Y. Daoyong, Estimation of multiple petrophysical parameters for hydrocarbon reservoirs with the ensemble-based technique. *Advances in Petroleum Exploration and Development* **4**(2), 1–17 (2012).
11. F. Aminzadeh and P. de Groot, *Neural Networks and Soft Computing Techniques with Application in the Oil Industry*, EAGE Publications, The Netherlands (2006).
12. Q-S. Xu and Y-Z. Liang, Monte Carlo cross validation. *Chemometr. Intell. Lab.*. **56**(1), 1–11 (2001).
13. M. Nikravesh and F. Aminzadeh, Past, present and future intelligent reservoir characterization trends. *J. Petrol. Sci. Eng.* **31**, 67–79 (2001).
14. S. Katz, G. Chilingar, and L. Khiliyk, Forecast and uncertainty analysis of production decline trends with bootstrap and Monte Carlo Modeling. *J. Sustain. Energy Eng.* **2**, 350–367 (2015).
15. N.E. Aase, P.A. Bjorkum, and P.H Nadeau, The effect of grain-coating microquartz on preservation of reservoir porosity. *AAPG Bull.* **80**(10), 1654–1673 (1996).
16. L. Breiman, Bagging predictors. *Machine Learning* **24**(2), 123–140 (1996).
17. L. Zhang and M. Poulton, Neural Network Inversion of EM39 induction log data, in *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*, W. Sandaham and M. Leggett (Eds.), pp. 231–247, Kluwer Academic Publishers, Dordrecht (2003).
18. R.K. Fruhwirth and S.P. Steinlechner, A systematic approach to the optimal design of feed forward neural networks applied to log-synthesis, *EAGE 66th Conference & Exhibition*, Expanded Abstracts (2004).
19. W.N. Venables and B.D. Ripley, *Modern Applied Statistics with S*, Fourth edition, Springer, The Netherlands (2002).

DOI: 10.7569/JSEE.2016.629519