IMPROVED PERMEABILITY PREDICTION USING

MULTIVARIATE ANALYSIS METHODS

A Thesis

by

JIANG XIE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2008

Major Subject: Petroleum Engineering

IMPROVED PERMEABILITY PREDICTION USING

MULTIVARIATE ANALYSIS METHODS

A Thesis

by

JIANG XIE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,   Akhil Datta-Gupta
Committee Members,   Yalchin Efendiev
                       Christine Ehlig-Economides
Head of Department,   Stephen A. Holditch

December 2008

Major Subject: Petroleum Engineering

ABSTRACT

Improved Permeability Prediction Using Multivariate Analysis Methods.

(December 2008)

Jiang Xie, B.S., University of Science and Technology of China

Chair of Advisory Committee: Dr. Akhil Datta-Gupta

Predicting rock permeability from well logs in uncored wells is an important task in reservoir characterization. Due to the high costs of coring and laboratory analysis, typically cores are acquired in only a few wells. Since most wells are logged, the common practice is to estimate permeability from logs using correlation equations developed from limited core data. Most commonly, permeability is estimated from various well logs using statistical regression.

For sandstones, often the logs of permeability can be correlated with porosity, but in carbonates the porosity permeability relationship tends to be much more complex and erratic. For this reason permeability prediction is a critical aspect of reservoir characterization in complex reservoirs such as carbonate reservoirs. In order to improve the permeability estimation in these reservoirs, several statistical regression techniques have already been tested in previous work to correlate permeability with different well logs. It has been shown that statistical regression for data correlation is quite promising in predicting complex reservoirs. But using all the possible well logs to predict permeability is not appropriate because the possibility of spurious correlation increases if you use more well logs. In statistics, variable selection is used to remove

unnecessary independent variables and give a better prediction. So we apply variable selection to the permeability prediction procedures in order to further improve permeability estimation.

We present three approaches to further improve reservoir permeability prediction based on well logs via data correlation and variable selection in this research. The first is a combination of stepwise algorithm with ACE technique. The second approach is the application of tree regression and cross-validation. The third is multivariate adaptive regression splines.

Three methods are tested and compared at two complex carbonate reservoirs in west Texas: Salt Creek Field Unit (SCFU) and North Robertson Unit (NRU). The result of SCFU shows that permeability prediction is improved by applying variable selection to non-parametric regression ACE while tree regression is unable to predict permeability because it can not preserve the continuity of permeability. In NRU, none of these three methods can predict permeability accurately. This is due to the high complexity of NRU reservoir and measurement accuracy. In this reservoir, high permeability is discrete from low permeability, which makes prediction even more difficult.

Permeability predictions based on well logs in complex carbonate reservoirs can be further improved by selecting appropriate well logs for data correlation. In comparing the relative predictive performance of the three regression methods, the stepwise with ACE method appears to outperform the other two methods.

DEDICATION

This thesis is dedicated to my parents for their love, care and inspiration.

ACKNOWLEDGEMENTS

I am going to remember these years of hard work with great pleasure. To all of you, I appreciate what you have done to help me in my scholastic and professional growth. I would like to thank you for providing me with a work environment that lends itself to creativity and productivity, without too many financial concerns. Not everyone is so fortunate. I know I still have much to learn, but with continued support and encouragement from people like you I know I can accomplish a great deal.

Thank you very much.

NOMENCLATURE

ACE   = alternating conditional expectations

AIC   = Akaike information criterion

BIC   = Bayes information criterion

DPHI   = density log derived porosity

DT   = acoustic transit time

EF   = electrofacies

GR   = gamma ray log

LLD   = lateral log deep

LLS   = lateral log shallow

MAE   = mean absolute error

MARS   = multivariate adaptive regression splines

MSE   = mean squared error

MSFL   = microspherically focused log

NPHI   = neutron log derived porosity

NRU   = North Robertson Unit

RHOB   = bulk density

RSS   = residual sum of squares

SCFU   = Salt Creek Field Unit

SP   = spontaneous potential

α = cost complexity factor

ρ = correlation coefficient

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 Literature Review

Predicting rock permeability from well logs in uncored wells is an important task in reservoir characterization. Due to the high costs of coring and laboratory analysis, permeability in most uncored wells is estimated using correlation equations developed from limited core data. Most commonly, permeability is estimated from various well logs using either an empirical relationship[1, 2] or some form of statistical regression: parametric or non-parametric.[1, 3-8] The empirical models may not be applicable in regions having different depositional environments without making adjustments to constants or exponents in the model. Also, significant uncertainty exists in the determination of irreducible water saturation and cementation factor in these models. Statistical regression has been proposed as a more versatile solution to the problem of permeability estimation. Conventional statistical regression is generally performed parametrically using multiple linear or nonlinear models that require *a priori* assumptions regarding functional forms.[3, 4]

Three important aspects in predicting permeability using statistical regression are:

1. Data correlation
2. Data partitioning
3. Variable selection

This thesis follows the style of *SPE Journal*.

1.1.1 Data Correlation

In recent years, non-parametric regression techniques such as Generalized Additive Model (GAM)[9], Alternating Conditional Expectation (ACE)[4-6, 10-12] and Neural Networks (NNET)[3] have been introduced to overcome the limitations of conventional multiple regression methods.[1, 5-8] Applications to complex carbonate reservoirs have shown great promise in handling many forms of heterogeneity in rock properties. However, significant difficulties remain in the identification of sharp local variations in reservoir properties caused by abrupt changes in the depositional environment. Another distinctive feature in carbonate reservoirs is the porosity-permeability mismatch, that is, low permeability in regions exhibiting high porosity and vice-versa. These features are extremely important from the point of view of fluid flow predictions, particularly for early breakthrough response along high permeability streaks.

Lee's work[10] improves permeability predictions in heterogeneous carbonate reservoirs through a combination of electrofacies characterization and non-parametric regression techniques. Three non-parametric approaches are considered: ACE, GAM, and NNET. In this study, tree regression[13, 14] and multivariate adaptive regression splines (MARS)[15-17] are considered and compared to Lee's results for a field case (Salt Creek Field Unit)

1.1.2 Data Partitioning

Another important aspect in data correlation for permeability predictions is data partitioning. A variety of approaches have been proposed to partition well log responses into distinct classes in order to improve permeability predictions. The simplest approach utilizes flow zones or reservoir layering.[6] Other approaches like lithofacies[18], Hydrologic Flow Unit (HFU)[7, 19-21] and electrofacies[10, 11, 22] are explored and widely used. Previous work by Lee[10] shows that electrofacies characterization with

the ACE algorithm, when compared against other data partitioning methods based on reservoir layering, lithofacies information and hydraulic flow units, gives a good permeability prediction. We will use electrofacies characterization as our data partitioning method and focus on other two aspects – data correlation and variable selection.

1.1.3 Variable Selection

A spurious correlation is defined as a sample correlation that is large in absolute value purely by chance.[23] Consider a well log as a possible predictor and the probability of observing the spurious correlation of this predictor is p_{sc}, we can calculate the probability of at least one spurious correlation when considering a set of k independent predictors, which is simply

$$1 - (1 - p_{sc})^k = \sum_{i=1}^{k} p_{sc}(1 - p_{sc})^{(i-1)}$$

We can see that the penalty for increasing the number of predictors considered from k-1 to k is

$$p_{sc}(1 - p_{sc})^{(k-1)}$$

As we use an increasing number of well logs, the chance of observing at least one spurious correlation increases. This is one of the reasons why we need to select appropriate variables for data correlation. Other possible problems with more well logs are:

1. Unnecessary predictor information will add noise to the estimation of other quantities that we are interested in.
2. Degrees of freedom will be wasted. Assume that n is sample size and k is number of independent variables, the degrees of freedom equals $d = n - k - 1$.

Degrees of freedom decreases as the number of independent variables increases.

Variable Selection is intended to select the "best" subset of predictors. Variable selection methods fall into two categories: best subset selection methods and stepwise regression methods. Best subset selection methods evaluate all the possible subsets of variables from a full model and identify the best reduced regression models based on some criterion. Evaluating all possible models is the most reasonable way to proceed in variable selection but it is computationally demanding, especially when one has a very large number of predictors. Stepwise algorithm is to evaluate the variables one at a time and look at a sequence of models.[24]

While stepwise algorithm is widely accepted for variable selection, pruning a regression tree may also be considered as a variable selection method. The difficulties of the stepwise algorithm and tree regression are the same: stopping criterion and optimal model selection. In the stepwise algorithm, a variety of criteria are available – AIC, BIC, adjusted R^2, Predicted Residual Sum of Squares (PRESS) and Mallow's C_p Statistic. We choose the AIC-criterion because the number of independent variables from AIC-criterion is a balance between model size and model regression error. In tree regression, cross-validation and AIC-criterion are two ways to select optimal tree size.[13, 14] We choose cross-validation for tree regression because AIC-criterion tends to give a larger optimal tree size and over-fit in prediction.

1.2 Objectives

The objective of this study is to further improve permeability prediction using well logs by variable selection and data correlation. First, stepwise regression is tested as a variable selection method. Then a method combining the stepwise algorithm and non-parametric regression is proposed and applied to a highly heterogeneous carbonate reservoir in the Permian Basin, west Texas: Salt Creek Field Unit (SCFU). Tree regression and multivariate adaptive regression splines (MARS) are performed for data correlation. Results are compared with non-parametric regression (ACE).

In statistics, stepwise regression is widely used for variable selection and model reduction. We start with stepwise regression as the linear regression model and apply it to North Robertson Unit (NRU) as a synthetic case in order to prove that stepwise algorithm can be used to improve permeability prediction. After that, a method combining the stepwise algorithm with non-parametric regression (ACE) is proposed and applied to Salt Creek Field Unit (SCFU). The result is compared with the ACE algorithm without variable selection.

In regression tree analysis, pruning might be considered as a method of variable selection[14]. Following this idea, we try to predict permeability using a regression tree and improve permeability estimation with a pruning tree. The key behind pruning trees is to select the optimal tree size. Cross-validation is a practical way to choose the optimal tree size. An AIC-criterion for pruning could also be used. Pruning with tree regression is applied to SCFU and NRU.

Another statistical regression method, multivariate adaptive regression splines (MARS), will be tested and applied to two field cases: SCFU and NRU. Results are compared with those from ACE algorithm.

1.3 Two Field Cases: SCFU and NRU

In this study, we are focusing on data correlation and variable selection aspects of permeability estimations. We use electrofacies characterization for data partitioning.

1.3.1 Salt Creek Field Unit (SCFU)

The Salt Creek field, located in Kent County, Texas, is a highly heterogeneous, shallow reef carbonate reservoir in the Basin shown in Figure 1.1. The reservoir lies on Strawn limestone and is overlain by the Cisco shale. The field consists of two major carbonate buildups: the South Main Body and the Northwest Extension. The buildup consists of grainstone shoals, bars and spits associated with a shelf environment. The average depth and thickness are 6300 ft and 170 ft respectively. The oil-water contact is located at -4425 ft subsea. Stratigraphically, the field can be subdivided into several vertical zones (labeled C5, C4, C3, C2b, C2a, C1b, and C1a) on the basis of paleontological picks and wireline correlations. Ten lithofacies have been identified from detailed core and thin-section studies: shale, siltstone, skeletal wackestone, algal wackestone, algal packstone, intraclastic packstone, skeletal/peloid packstone, peloid/skeletal mud lean packstone, peloid/sketetal grainstone, and ooid grainstone.[4, 25]
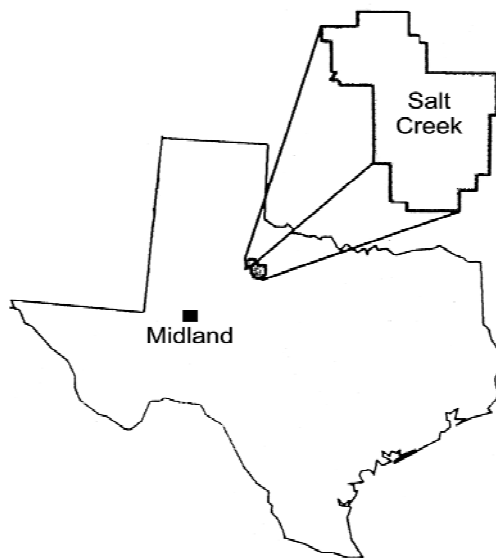
Figure 1.1 Location of Salt Creek Field Unit, Kent County, TX

In this study, we are focusing on the Northwest Extension area, which has 89 wells arranged in a 20-acre, five-spot pattern. The data presented in this analysis were collected from seven wells. We omitted two cored wells (G517 and G520) to verify our correlations with blind tests. 904 sample points from five wells are assigned as training data while 174 sample points from G517 and 183 sample points from G520 are assigned as prediction data. We have 10 well logs: caliper, spontaneous potential (SP), gamma ray (GR), three different resistivity (lateral log deep (LLD), lateral log shallow (LLS), and micro-spherically focused log(MSFL), acoustic transit-time (DT), neutron (NPHI), density (RHOB), and photoelectric logs (PEF). Seven well logs (GR, LLD, MSFL, DT, NPHI, RHOB and PEF) are selected out of ten well logs for data correlation and variable selection because of the similarity of the LLS and LLD responses, and also to account for the poor vertical resolution of SP and caliper logs.

Figure 1.2 Screenplot, A Barplot of the Variances of the Principal Components labeled by $\sum_{i=1}^{j} \lambda_i / trace(\Sigma)$

Principal Component Analysis is applied to obtain the principal components PCj (j=1,…,7) from the well log data after normalization. Figure 1.2 shows the screen plot, a bar plot of the variance of the principal components labeled by $\sum_{i=1}^{j} \lambda_i / trace(\boldsymbol{\Sigma})$ which often provides a convenient visual method of identifying the important components. Only 4 principal components explain around 90% variation of the whole data set. First principal component (PC1) appears to indicate porosity of the formation while second principal component (PC2) shows a stronger correlation with gamma ray readings. The eigenvectors of the covariance matrix provides coefficients of the principal components transformation. For example, PC1 and PC2 are given by

$$PC1 = 0.15GR + 0.36\log(LLD) + 0.42\log(MSFL)$$

$$-0.46DT + 0.47RHOB - 0.48NPHI + 0.1PEF$$

$$PC2 = 0.62GR - 0.31\log(LLD) - 0.14\log(MSFL)$$

$$-0.13DT + 0.09RHOB - 0.08NPHI - 0.68PEF$$

Model-based Cluster Analysis is used to define 8 distinct groups based on the unique characteristics of the well log measurements. In Figure 1.3, each cluster can be treated as an electrofacies that reflects the hydrologic, lithologic, and diagenetic characteristics. Qualitatively speaking, the first electrofacies group (EF1) indicates tight media with low gamma ray reading and the eighth electrofacies group (EF8) represents porous media with high gamma ray reading.
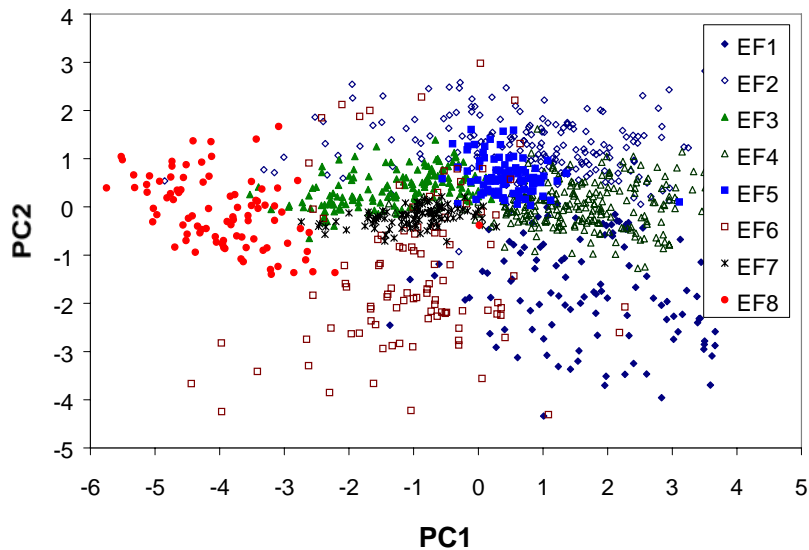
Figure 1.3 The Distribution of Electrofacies Data Plotted on the First Two Principal Components of Well Logs

1.3.2 North Robertson Unit (NRU)

North Robertson Unit is a heterogeneous shallow shelf carbonate reservoir in Gaines County, west Texas. The NRU is located in the northern part of the Central Basin Platform of the Basin shown in Figure 1.4. There are two main producing horizons, the Glorieta and Clearfork Formations, often referred to as the Upper and Lower Clearfork, between the correlative depths of approximately 5874 – 7440 feet. The reservoir interval is thick (gross interval = 1400 ft). More than 90% of the interval has uniform lithology (dolostone), but is characterized by a complex pore structure that results in extensive vertical layering.[7]

Reservoir sediments accumulated in the Clearfork area of the Central Basin Platform were deposited in shallow, agitated waters providing a favorable environment for carbonate generation. Early winnowing in this high-energy environment provided primary intergranular porosity in some carbonates. However, fluctuations in relative sea level resulted in cyclic deposition typical of many Permian stratigraphic units in west Texas. Also, leaching and dissolution of skeletal grains, as well as selective dolomitization, modified porosity distribution significantly. In addition, local fractures and other late diagenetic processes affected the reservoir quality of the deposits, and the end result is a very heterogeneous and complex carbonate.[7, 11]
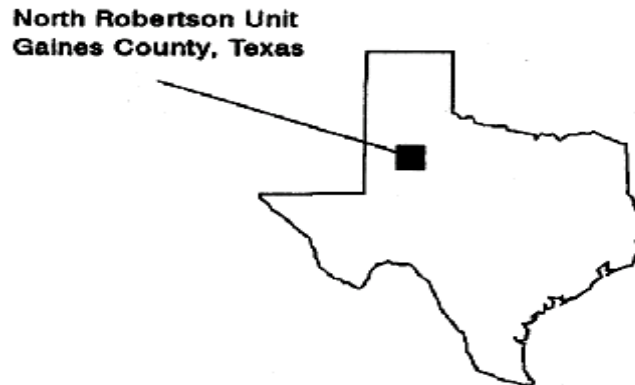
Figure 1.4 General Location of the North Robertson Unit
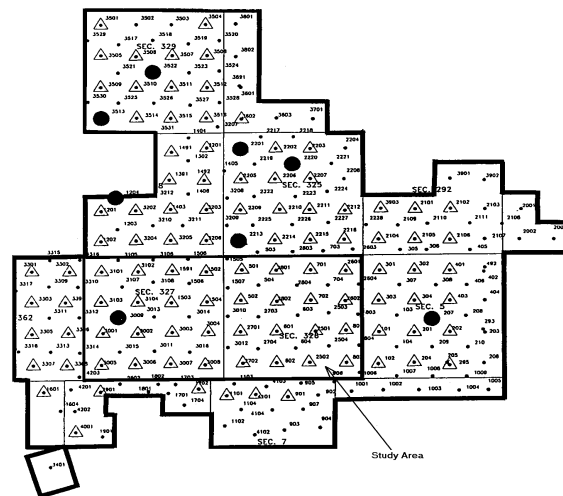
Figure 1.5 Location of Cored Wells in the North Robinson Unit

The data presented in this analysis were gathered from 6 cored wells and two wells (1509 and 3533) shown in Figure 1.5 are left out to verify our correlations using blind tests. 2630 sample points from six wells are assigned as training data while 712 sample points from W1509 and 851 sample points from W3533 are assigned as prediction data.

In this field, we have 9 well logs in different wells: caliper, spontaneous potential (SP), gamma ray (GR), two different resistivity logs (LLD and MSFL), acoustic transit-time or sonic (DT), neutron (NPHI), density (RHOB), and photoelectric (PEF). Only 6 well logs (GR, LLD, MSFL, DT, NPHI, and RHOB) are chosen because of their field-wide availability.

Electrofacies Characterization: Principal component analysis is used to summarize the data effectively and to reduce the dimensionality without any significant loss of information. This analysis is applied to obtain the principal components PCj (j=1,…,6) from the well log data after normalization.

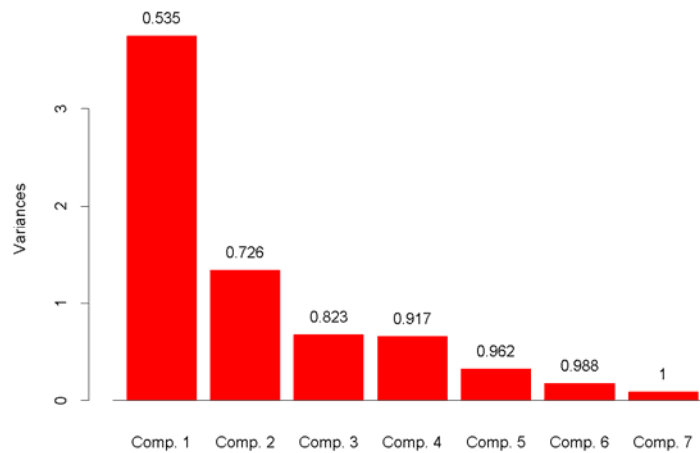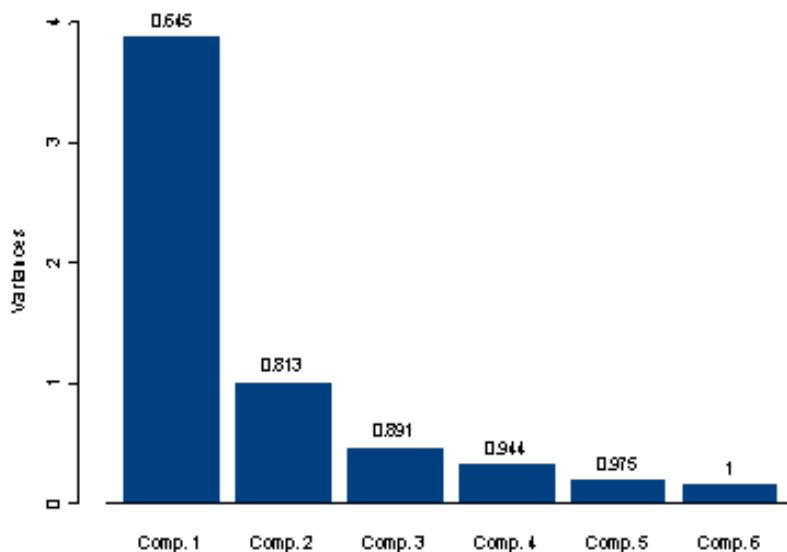Figure 1.6 A Barplot Representing the Variances Described by the PC's

Figure 1.6 shows the screen plot, a barplot of the variance of the principal

components labeled by

$$\sum_{i=1}^{l} \lambda_i / trace(\Sigma),$$

which often provides a convenient visual method of identifying the important principal components.

Only 3 principal components explain approximately 90% variation of the whole data set. First principal component (PC1) appears to indicate porosity of the formation while second principal component (PC2) shows a stronger correlation with gamma ray readings. The eigenvectors of the covariance matrix provides coefficients of the principal components transformation. For example, PC1 and PC2 are given by the following relationships

$PC1 = -0.16GR + 0.46\log(LLD) + 0.46\log(LLS)$

$\quad\quad -0.46DT + 0.41RHOB - 0.42NPHI$

$PC2 = 0.93GR + 0.13\log(LLD) + 0.20\log(LLS)$

$\quad\quad -0.21DT + 0.06RHOB - 0.16NPHI$

Next we apply model-based cluster analysis to define distinct groups based on the unique characteristics of the well log measurements. Each cluster can be treated as an electrofacies that reflects the hydrologic, lithologic, and diagenetic characteristics. The clustering algorithm indicates 9 distinct groups based on the well log measurements (Figure 1.7).
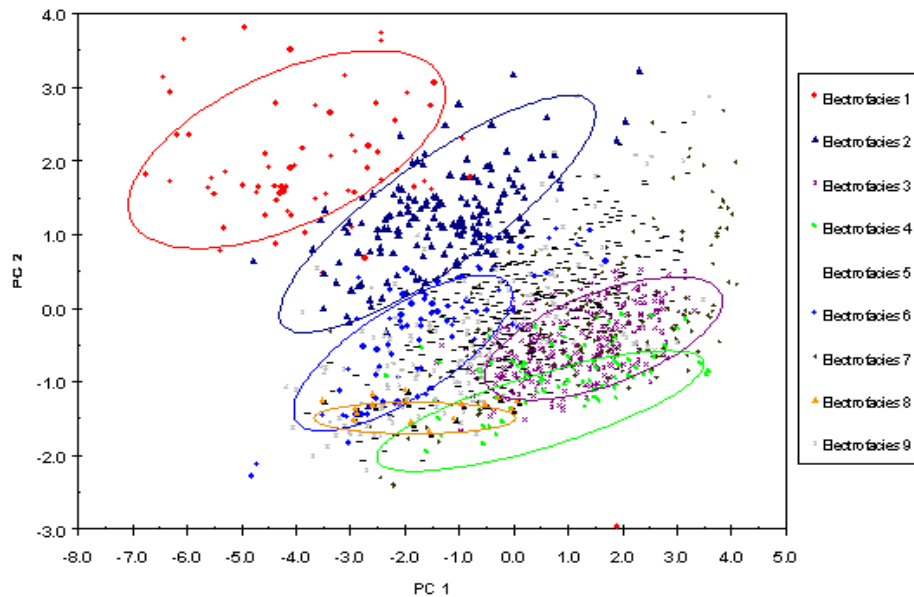
Figure 1.7 Electrofacies Characterization via Cluster Analysis

A qualitative interpretation of the clusters is aided by the relationship. For example, we can say that the first electrofacies group (EF1) represents porous media with high gamma ray reading and the ninth electrofacies group (EF9) represents tight media with low gamma ray reading.

1.4 Error Measurement

In statistics, the mean squared error (MSE) and the mean absolute error (MAE) are two quantities used to measure the difference between prediction and true value. As the names suggest, the mean absolute error is an average of the absolute errors and the mean squared error is the second moment of the error:

$$MSE = E\big((\hat{y}_i - y_i)^2\big) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

$$MAE = E\big(|\hat{y}_i - y_i|\big) = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

y_i is defined as true value, $\hat{y}_i$ as the prediction or estimator, and N is number of data.

After permeability prediction with statistics regression, MSE and MAE are calculated to measure the prediction errors and compare different regression methods. Another measurement of data correlation for our approach is the correlation coefficient between measured permeability and predicted permeability.

$$\rho = \frac{\mathrm{cov}(y, \hat{y})}{\sqrt{\mathrm{var}(y)}\sqrt{\mathrm{var}(\hat{y})}}$$

The correlation coefficient indicates the degree of linear dependence between the variables. The closer the coefficient is to ± 1, the stronger the correlation between the variables, which reflects our prediction accuracy.

1.5 Thesis Outline

In Chapter II, we discuss stepwise algorithm as a variable selection method to improve non-parametric regression. Stepwise regression with a linear model is applied to illustrate the availability of the stepwise algorithm. A method combining the stepwise algorithm and ACE is proposed and applied to SCFU and NRU.

In Chapter III, we apply tree regression to permeability prediction and utilize cross-validation to select the optimal tree for variable selection. The results of SCFU and NRU from tree regression are compared with results from ACE.

In Chapter IV, the multivariate adaptive regression splines (MARS) method is compared with ACE. We utilize MARS software from Salford systems to perform multivariate adaptive regression splines.

CHAPTER II

STEPWISE ALGORITHM

2.1 Introduction

In previous papers, electrofacies characterization and non-parametric regression ACE technique are proven to be successful in predicting permeability in highly heterogeneous carbonate reservoirs[10, 11]. In this chapter, we utilize the stepwise algorithm with AIC-criteria as a variable selection method to further improve permeability estimates. Stepwise regression is a kind of stepwise algorithm used in linear regression. We first demonstrate the availability of the stepwise algorithm by applying stepwise regression to North Robertson Unit (NRU). A method combining the stepwise algorithm and ACE is proposed and applied to field case Salt Creek Field Unit (SCFU). Results are compared with those from regression without variable selection.

2.2 Methodology

In this section, we discuss the non-parametric regression technique, stepwise procedure and stepwise algorithm. A new method combining the stepwise procedure with ACE is proposed for permeability estimation.

2.2.1 Response Transformation Models: ACE Algorithm[14, 26]

The response transformation models generalize the additive model by allowing for a transformation of the response variable y. The models have the following general form:

$$\theta(y) = \alpha + \sum_{l=1}^{p} \phi_l(x_l) + \varepsilon$$

The main motivation behind response transformation is that often a simple additive model may not be appropriate for $E(y \mid x_1, x_2, ..., x_p)$, but may be quite appropriate for $E\{\theta(y) \mid x_1, x_2, ..., x_p\}$. An example of such models is the Alternating Conditional Expectation (ACE) algorithm.

The ACE algorithm, originally proposed by Breiman and Friedman[26], is used here to derive the optimal transformations of the data to maximize the correlation between permeability and well logs. Let us say that we have a dependent random variable Y (for example, permeability) and independent random variables $x_1, x_2, ..., x_p$ (for example, well logs). We first define arbitrary mean-zero transformations $\theta(y), \phi_1(x_1), ..., \phi_2(x_p)$. Suppose that a regression of the transformed dependent variable on the sum of transformed independent variables (under the constraint, $E[\theta^2(y)] = 1$) results in the following error:

$$e^2(\theta, \phi_1, ..., \phi_p) = E\{[\theta(y) - \sum_{l=1}^{p} \phi_l(x_l)]\}^2$$

Then, transformations $\theta^*(y), \phi_1^*(x_1), ..., \phi_2^*(x_p)$ are said to be optimal for regression if they satisfy the following

$$e^{*2}(\theta^*, \phi_1^*, ..., \phi_p^*) = \min_{\theta, \phi_1, ..., \phi_p} e^2(\theta, \phi_1, ..., \phi_p)$$

More details on the procedure to derive the optimal transforms can be found.[6, 26]

Briefly, our approach proceeds as follows:

Develop optimal non-parametric transforms for permeability and well log variables based on cored wells: $\theta(y), \phi_1(x_1), ..., \phi_2(x_p)$;

For an uncored well, given a set of well log response $\{x_{li}, ..., x_{pi}\}$ first estimate the

corresponding transforms $\left\{\phi_1^*(x_{li}),...,\phi_1^*(x_{pi})\right\}$ from step 1;

1. Estimate the optimal transform for permeability using the following relationship

$$\theta^*(y_i) = \sum_{l=1}^{p} \phi_l^*(x_{li});$$

2. Finally, predict permeability through back transformation

$$y_i^{pre} = \theta^{*-1}\left[\sum_{l=1}^{p} \phi_l^*(x_{li})\right].$$

Thus, our calculation involves p forward transformations of $\left\{x_{li},...,x_{pi}\right\}$ to $\left\{\phi_1^*(x_{li}),...,\phi_1^*(x_{pi})\right\}$, and a backward transformation. By restricting the transformation of the response variable to be monotone, we can ensure that θ^* is invertible. Non-parametric implies that no functional form is assumed between the dependent and independent variables and the transformations are derived solely based on the data set. The optimal correlation is given by plotting the transformed dependent variable against the sum of the transformed independent variables.

Permeability is correlated with well logs within each electrofacies group using a non-parametric regression method. As mentioned before, the non-parametric regression allows us to correlate permeability with well logs without a priori assumptions regarding the functional form of the correlation. Specifically, a user-friendly software, GRACE (GRaphical ACE) is used for this purpose.

2.2.2 Stepwise Algorithm[24]

Stepwise Procedures include backward elimination and forward selection. Backward Elimination is the simplest of all variable selection procedures and can be easily implemented without special software. The algorithm is as follows:

Start with all the variables in the model;

Remove the variable with highest p-value greater than critical;

Refit the model and go to 2;

Stop when all p-values are less than critical.

The critical is sometimes called the "p-to-remove" and does not have to be 5%. If prediction performance is the goal, then a 15-20% cut-off may work best.

Forward selection just reverses the backward method:

Start with no variables in the model;

For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than critical;

Continue until no new variable can be added.

Stepwise regression is a combination of the stepwise procedure and linear regression. In stepwise regression, Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) are two commonly used criteria. In general,

$$AIC = -2\log - \text{likelihood} + 2p$$

while, $BIC = -2\log - \text{likelihood} + p\log n$

p is number of variables and n is number of data.

For a linear regression model, the -2log-likelihood is known as the deviance, $n\log(RSS/n)$. In our study, we choose to minimize AIC in order to find the optimal model. Larger models will fit better and so have smaller RSS but use more parameters. Thus the best choice of model will balance fit with model size.

2.2.3 Combining ACE with Stepwise Procedure

A new method combining the non-parametric regression technique (ACE) with the stepwise procedure is proposed. The general idea is to replace linear regression in stepwise regression with the ACE algorithm. We still utilize AIC criteria for model selection and use both backward elimination and forward addition in our stepwise procedure. The procedure is the following:

Fit the training data with the full model (all well logs) using ACE;

Calculate AIC criteria;

Single term deletions and additions,

Non-parametric regression (ACE) with all new models,

Calculate AIC criteria of new models;

If there is a smaller AIC value, select that model with least AIC value and repeat step2. If no model has smaller AIC than the initial model, stop the stepwise procedure and select initial model as the optimal model.

When we combine the stepwise algorithm with ACE technique, the basic procedure follows the stepwise procedure while the ACE technique is only used for model fitting. In stepwise regression, the basic procedure is the same but linear regression is used to fit the model.

2.3 Field Example: North Robertson Unit

Stepwise regression is used for a field case to demonstrate how the stepwise procedure is performed. We will show that stepwise regression on North Robertson Unit improves the initial linear regression result.

Here, electrofacies VII is used as an example to show how stepwise regression works. Starting with the full model, which means that all six well logs (GR, DT, log10(LLD), log10(MSFL), NPHI and RHOB) are used, we follow stepwise procedure of backward elimination and forward addition. As each well log is added or eliminated, a linear regression is run and the AIC-criterion is calculated to compare with the optimal model from the last step until no more well logs can be added or eliminated. The following shows the stepwise procedure and stepwise regression result of electrofacies VII from an S-plus report.

Table 2.1 Stepwise Regression Procedures and S-plus Report of Electrofacies VII, NRU

Explanation	S-plus Report

Starting full model with six well logs	```*** Stepwise Regression ***```
	
	``` *** Stepwise Model Comparisons ***```
	Start: AIC= 273.0431
	
	lnkcore ~ GR + NPHI + RHOB + DT + log.LLD. +
	
	log.MSFL.

Step 1
 	
Backward elimination: (no forward addition because no well logs could be added)
	Single term deletions
	```Model:```
	```lnkcore ~ GR + NPHI + RHOB + DT + log.LLD. + log.MSFL.```
	
	```scale: 3.033812```

Table 2.1 Continued

Explanation	S-plus Report

Calculate regression models with only one well log eliminated.
 	
Choose the best model with least AIC value.
 	
Here, well log GR is eliminated.
	``
	       Df Sum of Sq RSS Cp
	  <none> 230.5697 273.0431
	    GR 1 1.43069 232.0004 268.4061
	   NPHI 1 11.51604 242.0857 278.4915
	   RHOB 1 2.78198 233.3517 269.7574
	     DT 1 1.96724 232.5369 268.9427
	 log.LLD. 1 8.73986 239.3096 275.7153
	log.MSFL. 1 6.83604 237.4057 273.8115
	
	Step: AIC= 268.4061
	``
	$$lnkcore \sim NPHI + RHOB + DT + log.LLD. + $$
	$$log.MSFL.$$

Step 2
 	
Backward elimination & forward addition:
 	
Calculate regression models with only one well log eliminated or added.
 	
Choose the best model.
 	
Here, well log RHOB is eliminated.
	``
	Single term deletions
	Model:
	lnkcore ~ NPHI + RHOB + DT + log.LLD. + log.MSFL.
	scale: 3.033812
	
	       Df Sum of Sq RSS Cp
	  <none> 232.0004 268.4061
	   NPHI 1 10.09090 242.0913 272.4294
	   RHOB 1 2.62001 234.6204 264.9585
	     DT 1 2.90834 234.9087 265.2468
	 log.LLD. 1 7.31938 239.3198 269.6579
	log.MSFL. 1 8.01270 240.0131 270.3512
	
	Single term additions
	Model:
	lnkcore ~ NPHI + RHOB + DT + log.LLD. + log.MSFL.
	scale: 3.033812
	       Df Sum of Sq RSS Cp
	<none> 232.0004 268.4061
	   GR 1 1.430695 230.5697 273.0431
	
	Step: AIC= 264.9585
	 lnkcore ~ NPHI + DT + log.LLD. + log.MSFL.
	``

Table 2.1 Continued

Explanation	S-plus Report

Step 3
 	
Backward elimination & forward addition
 	
Well log DT is eliminated
	Single term deletions
	Model:
	lnkcore ~ NPHI + DT + log.LLD. + log.MSFL.
	scale: 3.033812
	
	      Df Sum of Sq RSS Cp
	  <none> 234.6204 264.9585
	    NPHI 1 10.41612 245.0365 269.3070
	    DT 1 1.03523 235.6556 259.9261
	 log.LLD. 1 9.57489 244.1953 268.4658
	log.MSFL. 1 12.19701 246.8174 271.0879
	
	Single term additions
	Model:
	lnkcore ~ NPHI + DT + log.LLD. + log.MSFL.
	scale: 3.033812
	
	      Df Sum of Sq RSS Cp
	<none> 234.6204 264.9585
	   GR 1 1.268725 233.3517 269.7574
	 RHOB 1 2.620006 232.0004 268.4061
	
	Step: AIC= 259.9261
	 lnkcore ~ NPHI + log.LLD. + log.MSFL.

Step 4
 	
Backward elimination & forward addition.
 	
With single term deletion and addition, no other regression model is better than initial model in term of AIC criteria.
	Single term deletions
	Model:
	lnkcore ~ NPHI + log.LLD. + log.MSFL.
	scale: 3.033812
	
	      Df Sum of Sq RSS Cp
	  <none> 235.6556 259.9261
	    NPHI 1 15.3416 250.9972 269.2001
	 log.LLD. 1 46.7732 282.4288 300.6317
	log.MSFL. 1 13.6257 249.2813 267.4842
	
	Single term additions

Table 2.1 Continued

Explanation	S-plus Report
| Stepwise regression is stopped. Variables (NPHI, log10_LLD, log10_MSFL) are selected. | ```
Model:
lnkcore ~ NPHI + log.LLD. + log.MSFL.
scale:  3.033812


     Df Sum of Sq      RSS       Cp
<none>              235.6556 259.9261
   GR  1  1.857994 233.7976 264.1357
 RHOB  1  0.746895 234.9087 265.2468
   DT  1  1.035225 234.6204 264.9585
``` |
| Linear regression of optimal model is performed. | ```
   *** Linear Model ***


Call: lm(formula = lnkcore ~ NPHI + log.LLD. +
log.MSFL., data = NRU.EF7, na.action
     = na.exclude)
Residuals:
   Min    1Q Median    3Q   Max
 -3.496 -1.119 -0.1982 0.6347 4.576


Coefficients:
             Value Std. Error  t value Pr(>|t|)
(Intercept) -6.0195   1.0103   -5.9583   0.0000
     NPHI 12.4082   5.4714    2.2678   0.0261
  log.LLD.  1.6695   0.4216    3.9598   0.0002
 log.MSFL. -0.7314   0.3422   -2.1372   0.0357


Residual standard error: 1.727 on 79 degrees of freedom
Multiple R-Squared: 0.1834
F-statistic: 5.915 on 3 and 79 degrees of freedom, the
p-value is 0.001078
``` |

After stepwise regression, an optimal model is selected, called the stepwise model. We perform regression and prediction with both the stepwise model and full model (all six well logs are used). Regression error and prediction errors of electrofacies VII are

shown in Table 2.2. From MSE and MAE of electrofacies VII, we can see that the stepwise procedure selects a generalized model and gives better predictions of well 1509 and 3533 compared to the full model. Thus, we can consider the stepwise algorithm as a variable selection and model reduction procedure.

Table 2.2 Regression and Prediction Errors of Full Linear model and Stepwise Linear Model of Electrofacies VII, NRU

	Error	Full model	Stepwise model
Regression	MSE	2.7780	2.8392
	MAE	1.2667	1.3022
Prediction W1509	MSE	0.8885	0.6058
	MAE	0.7355	0.6342
Prediction W3353	MSE	1.4308	1.0176
	MAE	0.9763	0.8078

Table 2.3 Variables Selected of All Electrofacies, NRU

Electrofacies	Variables selected
I	No variables
II	RHOB, log10_LLD, log10_MSFL
III	NPHI, RHOB, log10_LLD
IV	NPHI, DT, log10_LLD
V	GR, RHOB
VI	NPHI
VII	NPHI, log10_LLD, log10_MSFL
VIII	NPHI, RHOB, DT, log10_MSFL
IX	GR, RHOB

Table 2.3 shows the selected variables of each electrofacies. Different electrofacies represent different geologic zones or regions and can be characterized by specific variables. In electrofacies I, no variable is selected, which means that constant fitting is better than all other linear fits. This is the limitation of the linear regression model.

Table 2.4 Regression and Prediction Errors of All Electrofacies, NRU

	Error	Full model	Stepwise model
Regression	MSE	2.7724	2.8776
	MAE	1.2864	1.3225
Prediction W1509	MSE	1.3535	1.3376
	MAE	0.8068	0.7661
Prediction W3353	MSE	2.8961	2.5827
	MAE	1.0985	1.0712

The mean squared errors and mean absolute errors are compared in Table 2.4. Errors show us the improvement of the stepwise model in predicting permeability. The stepwise model gives larger regression error in training data because the stepwise model is a more general model for all data sets and the full model is more specific for this training data set. Table 2.4 shows that the stepwise procedure can be used as a variable selection method to improve permeability estimate.

Figure 2.1 Measured Permeability vs. Predicted, W1509 NRU, Stepwise Regression

From the comparison of measured permeability and predicted permeability in Figure 2.1 and Figure 2.2, we can not see a good permeability prediction. Figure 2.2 indicates that stepwise regression is unable to predict the general variation of W1509 and correlation coefficient $\rho = 0.1950$ shows low linear dependency. This may be a result of the limitation of the stepwise regression method or the permeability data from

North Robertson Unit (NRU).

Figure 2.2 Measured and Predicted Permeability vs. Depth, W1509 NRU, Stepwise Regression

2.4 Field Application: Salt Creek Field Unit

The new method combining the stepwise algorithm with ACE technique is tested with data from Salt Creek Field Unit (SCFU). After electrofacies characterization, training data are grouped into eight electrofacies and data from G517 and G520 are used for a blind test and prediction.

We first apply the new method to electrofacies VII as a simple example: Table 2.5 summarizes the procedure of this method. In step1, the initial model is the full model with all seven well logs. After single term deletions and ACE regression, RSS (residual sum of squares), AIC and BIC criteria are calculated. The model with well log DT deleted is selected as the initial model for the next step.

In step2, we have the same single term deletions and additions and non-parametric regression. After calculation of the AIC citeron, we found that no other regression model is better than the initial model, thus the stepwise procedure is stopped and the optimal model is selected with six well logs (GR, log10_LLD, log10_MSFL, NPHI, PEF and RHOB). The optimal model will be applied to ACE. The following figures show the result of software GRACE.

Table 2.5 Stepwise Procedure of Electroafacies VII, SCFU

Step 1	RSS	AIC	BIC	
7 logs	3961.819	308.6657	155.3264	
- DT	**3842.11**	**305.5331**	**152.2643**	**-DT**
- GR	4152.886	308.4043	155.1355	
- log10(LLD)	5018.874	315.3962	162.1274	
- log10(MSFL)	4276.268	309.4852	156.2164	
- NPHI	4244.215	309.2075	155.9387	
- PEF	4175.475	308.6047	155.3359	
- RHOB	4335.877	309.9962	156.7274	
Step 2	RSS	AIC	BIC	
6 logs	**3842.11**	**305.5331**	**152.2643**	**Stop**
- GR	4058.689	305.5575	152.3593	
- log10(LLD)	5972.486	319.8179	166.6197	
- log10(MSFL)	5274.891	315.2328	162.0346	
- NPHI	4102.583	305.9546	152.7564	
- PEF	4279.182	307.5103	154.3122	
- RHOB	4238.882	307.161	153.9629	

Optimal Transform

$y = -1.5147E-03x^3 + 3.8003E-02x^2 - 2.0346E-01x - 2.5609E-01$
$R^2 = 9.5336E-01$

Figure 2.3 Optimal Transform of GR in Electrofacies VII, SCFU, ACE Algorithm

Optimal Transform

$y = -1.4166E+01x^3 + 8.4645E+01x^2 - 1.6950E+02x + 1.1366E+02$
$R^2 = 9.2401E-01$

Figure 2.4 Optimal Transform of LLD in Electrofacies VII, SCFU, ACE Algorithm

Optimal Transform

Figure 2.5 Optimal Transform of MSFL in Electrofacies VII, SCFU, ACE Algorithm

Optimal Transform

Figure 2.6 Optimal Transform of NPHI in Electrofacies VII, SCFU, ACE Algorithm

Optimal Transform

Figure 2.7 Optimal Transform of PEF in Electrofacies VII, SCFU, ACE Algorithm

Optimal Transform

Figure 2.8 Optimal Transform of RHOB in Electrofacies VII, SCFU, ACE Algorithm

The general idea of non-parametric regression ACE algorithm is to transform independent and dependent variables into optimal space with maximized linear correlation. We utilized the GRACE software developed by Peter Valko and Akhil Datta-Gupta to perform the non-parametric regression ACE technique. Figure 2.8, Figure 2.4, Figure 2.5, Figure 2.6, Figure 2.7 and Figure 2.8 give the relations between independent variables and their transformed values. We fit those curves with polynomials in order to get simple correlations. In optimal space, we build a linear correlation between ln(Kg)_Tr and Sum_Tr (sum of all variables_Tr) in Figure 2.9. The correlation coefficient is 0.7089 for this case.

Optimal Regression, Correl:0.7089

Figure 2.9 Optimal Regression ln(Kg)_Tr vs. Sum_Tr_Indep of Electrofacies VII, SCFU, ACE Algorithm

Figure 2.10 Optimal Inverse Transform of Dependent Variable of Electrofacies VII, SCFU, ACE Algorithm

Fitted   Stdev = 1.1785

GR_Tr= -1.5147E-03x3 + 3.8003E-02x2 - 2.0346E-01x - 2.5609E-01

log10(LLD)_Tr= -1.4166E+01x3 + 8.4645E+01x2 - 1.6950E+02x + 1.1366E+02

log10(MSFL)_Tr= 2.5858E+01x3 - 1.3513E+02x2 + 2.3181E+02x - 1.3067E+02

NPHI_Tr= -5.8188E-01x2 - 1.2062E+01x + 1.1224E+00

PEF_Tr= -6.0319E-01x2 + 6.9066E+00x - 1.9470E+01

RHOB_Tr= -5.1454E+01x2 + 2.5098E+02x - 3.0509E+02

ln(Kg)= 2.2505E-01 SumTr2 + 1.7926E+00 SumTr + 3.4391E-01

Figure 2.11 Cross plot of Fitted vs. Measured Permeability in Electrofacies VII

Figure 2.10 is an inverse transform of ln(Kg) from transformed space. Figure 2.11 is the final fitting vs. measured log-permeability. With polynomial equations from Figure 2.11, we are able to predict the new data set (G517 and G520) and calculate errors (MSE and MAE).

Table 2.6 shows the improved permeability estimate in terms of reduced errors. Compared to the full model, the optimal stepwise is a generalized model with better permeability prediction ability.

Table 2.6 Regression and Prediction Errors of Electrofacies VII, SCFU

	Error	Full model	Stepwise model
Regression	MSE	1.2202	1.3895
	MAE	0.8697	0.9234
Prediction G517	MSE	2.4793	1.0917
	MAE	1.1647	0.8439
Prediction G520	MSE	1.7861	1.5227
	MAE	1.3361	0.7525

In order to apply the new method to all the other electrofacies, we select variables and obtain the optimal model for each electrofacies in Table 2.7 and then calculate the mean squared error and the mean absolute error in Table 2.8.

Table 2.7 Variables Selected of All Electrofacies, SCFU

Electrofacies	Variables Selected
I	DT, GR, log10_MSFL, PEF
II	GR, log10_LLD, log10_MSFL
III	DT, GR, log10_LLD, PEF, RHOB
IV	DT, GR, log10_LLD, log10_MSFL, RHOB
V	DT, GR, log10_MSFL, NPHI, RHOB
VI	DT, GR, log10_LLD
VII	GR, log10_LLD, log10_MSFL, NPHI, PEF, RHOB
VIII	DT, GR, log10_LLD, log10_MSFL, NPHI, PEF, RHOB

Table 2.8 Regression and Prediction Errors of All Electrofacies, SCFU

	Error	Full model	Stepwise model
Regression	MSE	1.58	1.6850
	MAE	0.97	1.1033
Prediction G517	MSE	2.25	1.9728
	MAE	1.15	1.0682
Prediction G520	MSE	1.74	1.8127
	MAE	1.04	1.0592

Finally, we plot measured permeability vs. predicted in Figure 2.12 and plot measured and predicted permeability vs. depth in Figure 2.13. Figure 2.12 indicates good linearity between measured permeability and predicted permeability with correlation coefficient = 0.6967. From Figure 2.13, we see that in general prediction is able to define the main feature of measured permeability vs. depth, but less successful in reproducing detailed extreme permeability. Permeability prediction is good considering that this is a highly heterogeneous carbonate reservoir.

Figure 2.12 Measured Permeability vs. Predicted Permeability, G517 SCFU, Stepwise ACE Algorithm

Figure 2.13 Measured and Predicted Permeability vs. Depth, G517 SCFU, Stepwise ACE Algorithm

2.5 Summary

In this chapter, we discussed the stepwise algorithm, stepwise regression and proposed a new method combining the stepwise procedure with ACE algorithm to predict permeability. Our results show that the proposed method can improve permeability prediction under highly heterogeneous reservoir environments. Statistical software S-Plus is used for stepwise regression. Computer program GRACE (data correlation with ACE model developed by Peter Valko and Akhil Datta-Gupta) and manual stepwise procedure is performed with our new proposed method.

The following specific conclusions can be drawn based on this chapter:

1. The stepwise algorithm in conjunction with non-parametric regression (ACE) technique shows potential for permeability predictions in complex carbonate reservoirs.

2. An examination of the errors for uncored wells indicates that the stepwise algorithm potentially improves permeability estimation with both linear model and non-parametric regression model.

3. Variable selection and model reduction can be used further in permeability prediction.

4. Non-parametric regression techniques provide a practical tool for estimating permeability in heterogeneous carbonate reservoirs.

CHAPTER III

TREE REGRESSION

3.1 Introduction

In this chapter, we discus regression tree and cross validation and their applications in data correlation and variable selection. The regression tree method is used to generate an initial tree model and cross-validation is used to find the optimal tree size. Cross validation and pruning tree together can be considered as a way of variable selection. After pruning, the optimal tree model is used to fit training data and predict the blind test data. Tree regression method is applied to both SCFU and NRU. Results are compared to non-parametric regression ACE.

3.2 Methodology

3.2.1 Regression Tree[13, 14]

A regression tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting them further on each of the branches. Initially all of the records in a training set (the pre-classified records that are used to determine the structure of the tree) are lumped in one big box. The algorithm then tries breaking up the data, using every possible binary split on every field. The algorithm chooses the split that partitions the data into two parts such that the sum of the squared deviations from the mean in the separate parts is minimized. This splitting or partitioning is then applied to each of the new branches. The process continues until each node reaches a user-specified minimum node size and becomes a

terminal node. (If the sum of squared deviations from the mean in a node is zero, then that node is considered a terminal node even if it has not reached the minimum size.)

3.2.2 Cross-validation and Pruning

Since the tree is grown from the training data set, when it has reached full structure it usually suffers from over-fitting (i.e. it is "explaining" random elements of the training data that are not likely to be features of the larger population of data). This results in poor performance on real life data. Therefore, it has to be pruned using the validation data set. Cross-validation[13] is a very practical way to choose optimal tree size. It requires splitting data into k subsets, with a default choice of $k=10$. Each of the k subsets of the data is left out in turn, the model is fitted to the remaining data, and the results used to predict the outcome for the subset left out are compared with the initial tree regression results. In a regression model, prediction error is usually taken as the sum of differences between observed and predicted, that is, the criterion is the same as that used for the splitting rule.

Pruning can also be considered as a method of variable selection. AIC penalizes minus twice log-likelihood by twice the number of parameters. For regression tree, we can take $\alpha = 2\hat{\sigma}^2$. One way to select $\hat{\sigma}^2$ would be from the fit of the full tree model. But AIC-criterion tend to over-fit and choose larger constants[14] because we usually underestimate $\hat{\sigma}^2$ from the full tree model.

3.3 Tree Regression Procedures

This section lists tree regression and cross-validation procedures and S-plus commands as an illustrative workflow. S-plus is an interface-based statistical software package; however, the command line can also be used to run the program. In the following

procedure, we find the optimal tree and make a prediction for each electrofacies. After predictions, mean squared error (MSE) and mean absolute error (MAE) are calculated.

Procedure:

1. Generate full tree model with mincut = 2, minsize = 4, mindev = 10^{-6} and plot full tree,

   tree.full <- tree(predictor ~ ., data = electrofacies, mindev = 10^{-6}, mincut = 2, minsize = 4);

   plot.tree(full.tree, "uniform"); text.tree(full.tree);

2. Cross-validation to find optimal tree size,

   tree.cv <- cv.tree(tree.full, , prune.tree);

   for (i in 2:5) tree.cv$dev <- tree.cv$dev + cv.tree(tree.full, , prune.tree)$dev;

   tree.cv$dev <- tree.cv$dev/5;

   plot(tree.cv);

3. Prune full tree model to optimal tree size and plot optimal tree,

   tree.prune <- prune.tree(tree.full, best = optimal size);

4. Predict new data set with optimal tree model,

   tree.predict <- predict.tree(tree.prune, newdata = list(), type="vector");

3.4 Field Application: Salt Creek Field Unit

After electrofacies characterization, the training data set is divided into eight electrofacies. In each of electrofacies, tree regression is performed according to the workflow. Now we take electrofacies I as an example to demonstrate this regression tree method.

Figure 3.1 shows a regression tree for data correlation between the logarithm of permeability and well logs in electrofacies I. The numbers in each terminal node represent the final fitted logarithm of permeability for this regression tree model. This initial tree model has 47 terminal nodes and a residual mean deviance of 0.562. These attributes represent the complexity of the tree and its accuracy. The initial regression tree model is used to fit the training data itself and predict logarithm permeability in electrofacies I of the blind wells G517 and G520.

Figure 3.1 Initial Regression Tree Model for Electrofacies I at SCFU

Using a cross-validation chart, we select the optimal tree size and prune the initial tree to this size. Figure 3.2 shows the cross-validation chart of electrofacies one and Figure 3.3 shows the optimal regression tree after pruning with an optimal tree size of 5. The number of terminal nodes is reduced from 47 to 5 with an increase in residual mean deviance of 0.562 to 1.799. However, the accuracy of the prediction should be improved by cross-validation and pruning.

Figure 3.2 Cross-validation of Electrofacies I at SCFU

Figure 3.3 Optimal Regression Tree Model for Electrofacies I at SCFU

Next, we use the optimal regression tree model to fit training data and predict blind well G517 and G520. Table 3.1 is a comparison of regression error and prediction error of the initial tree model and optimal tree model for electrofacies I. The MSE and MAE result proves that a simpler tree structure leads to a better predictive ability because unseen samples are less sensitive to the statistical irregularities and idiosyncrasies of the training data.

Table 3.1 Comparison of Regression and Prediction Errors of Initial Tree and Optimal Tree Model of Electrofacies I, SCFU

	Error	Initial model	Optimal model
Regression error	MSE	0.3241	1.718
	MAE	0.4071	1.076
Prediction error G517	MSE	No data in EF1	
	MAE		
Prediction error G520	MSE	5.6609	4.2986
	MAE	1.8817	1.7656

The advantage of pruning and cross-validation is clear from Table 3.1. Moreover, pruning tree and cross-validation can be considered as a variable selection method to improve permeability prediction. In electrofacies I, only GR, RHOB and LLD are selected for data correlation out of seven well logs.

This above-stated workflow is repeated for the other seven electrofacies and the data from all electrofacies are combined. Table 3.2 shows the optimal tree size and, variables selected for all electrofacies and Table 3.3 shows regression error and prediction error of the initial and optimal models for all electrofacies. As expected, prediction errors of blind well G517 and G520 are all reduced and regression errors increase after pruning.

Table 3.2 Optimal Tree Size and Variables Selected for All Electrofacies, SCFU

Electrofacies	Optimal tree size	Variables selected
I	5	GR, RHOB, LLD
II	6	RHOB, DT, PEF, MSFL, NPHI
III	3	RHOB
IV	4	GR, RHOB, LLD
V	2	LLD
VI	12	GR, RHOB, DT, PEF, MSFL, LLD, NPHI
VII	2	DT
VIII	3	GR, DT

Table 3.3 Comparison of Regression and Prediction Errors of Initial and Optimal Tree Model for All Electrofacies, SCFU

	Error	Initial model	Optimal model
Regression error	MSE	0.1995	1.4362
	MAE	0.3031	0.9454
Prediction error G517	MSE	3.2761	2.2546
	MAE	1.4278	1.1798
Prediction error G520	MSE	4.1523	3.0595
	MAE	1.6867	1.4078

Figure 3.4 Measured Permeability vs. Predicted Permeability, G517 SCFU, Tree Regression

Figure 3.5 Measured and Predicted Permeability vs. Depth, G517 SCFU, Tree Regression

Figure 3.4 shows a cross plot of predicted permeability vs. measured permeability of blind well G517. Figure 3.5 gives the plot of measured and predicted permeability vs. depth for well G517. From these two figures we observe that tree regression captures the general feature of measured permeability, but the limitation of tree regression is also obvious in that the same value tends to be assigned to different measured permeability because the average permeability prediction is given in each node and variation inside node is ignored. However, from the correlation coefficient

point of view, a correlation with $\rho = 0.6757$ is acceptable for permeability prediction.

Comparing the mean squared error and the mean absolute error of tree regression model to the ACE model in Table 3.4, we can say that tree regression with pruning gives competitive prediction errors to ACE. Moreover, tree regression tends to underestimate extreme permeability while the ACE algorithm sometimes leads to the wrong direction.

Table 3.4 Comparison of Regression and Prediction Errors of Tree Regression and ACE for All Electrofacies, SCFU

	Error	Tree regression with pruning	ACE[25]
Regression error	MSE	1.4362	1.58
	MAE	0.9454	0.97
Prediction error G517	MSE	2.2546	2.25
	MAE	1.1798	1.15
Prediction error G520	MSE	3.0595	1.74
	MAE	1.4078	1.04

3.5 Field Application: North Robertson Unit

Following the same workflow, we calculate prediction and regression of all electrofacies. Table 3.5 summarizes regression error and prediction error of the initial tree model and optimal models. The optimal model greatly reduces prediction errors and also increases regression errors, which indicates the improved permeability prediction by cross-validation and pruning tree.

Table 3.5 Comparison of Regression and Prediction Errors of Initial and Optimal Tree Model for All Electrofacies, NRU

	Error	Initial model	Optimal model
Regression error	MSE	0.4971	2.5210
	MAE	0.4348	1.2288
Prediction error W1509	MSE	3.3367	1.5286
	MAE	1.4431	0.8500
Prediction error W3533	MSE	4.9582	2.8074
	MAE	1.7247	1.1510

Table 3.6 Comparison of Regression and Prediction Errors of Tree Regression and
ACE for All Electrofacies, NRU

	Error	Tree regression with pruning	Stepwise
Regression error	MSE	2.5210	2.8776
	MAE	1.2288	1.3225
Prediction error W1509	MSE	1.5286	1.3376
	MAE	0.8500	0.7661
Prediction error W3533	MSE	2.8074	2.5827
	MAE	1.1510	1.0712

Comparing errors from the tree regression and stepwise regression method in Table 3.6, we can see that tree regression is unable to predict permeability well. However, we believe that this is not because of the limitation of the tree regression method.

Figure 3.6 Measured Permeability vs. Predicted Permeability, W1509 NRU, Tree Regression

Figure 3.6 and Figure 3.7 compare the measured permeability and predicted permeability in two different ways. Obviously, we can see the lack of linear dependence in Figure 3.6 and bad fitting in Figure 3.7. The correlation coefficient $\rho = 0.1339$ also shows the same result.

Figure 3.7 Measured and Predicted Permeability vs. Depth, W1509 NRU, Tree Regression

3.6 Summary

In this chapter, we applied tree regression method to our data correlation and used cross-validation and pruning to generalize a tree model for further improvement of permeability estimation. We apply this method to two field cases: Salt Creek Field Unit (SCFU) and North Robertson Unit (NRU). Result from SCFU shows that tree regression with pruning failed to predict permeability in carbonate reservoirs because it is unable to preserve the continuity of permeability. The reason is that in tree regression limited terminal nodes give limited predicted permeability and variation inside each node is ignored.

Some conclusions are drawn:

1. Tree regression fails to predict permeability in carbonate reservoirs because it fails to capture the continuity of permeability.

2. Cross-validation is a more efficient method of the variable selection process. Combined with a pruning tree, cross-validation greatly reduces prediction errors and gives an improved permeability estimate.

CHAPTER IV

MULTIVARIATE ADAPTIVE REGRESSION SPLINES

4.1 Introduction

Compared to the regression tree method, the multivariate adaptive regression splines (MARS) method is considered as a generalization of recursive partitioning regression. In this chapter, we apply MARS to permeability estimation and compare it with the ACE algorithm and regression tree. The MARS software from Salford Systems is used to build the regression model and predict permeability.

4.2 Methodology

The multivariate adaptive regression splines method[15-17] is a generalization of recursive partitioning regression. The MARS model is in the form of

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$$

The model is a weighted sum of basis functions. Each c_i is a constant coefficient. The Mth basis function is like this:

$$B_m(x) = \prod_{k=1}^{Km} H\left[s_{km} \cdot \left(x_v(k,m) - t_{km}\right)\right].$$

H[η] is a Hinge function,

$$H[\eta] = \begin{cases} 1 & \text{if } \eta \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

MARS builds a model in two phases: the forward and the backward pass. This two stage approach is the same as that used by recursive partitioning trees. Forward

stepwise knot placement is used to generate all possible models within limitation, then backward stepwise elimination with generalized cross validation is performed to select the best model.

The MARS software from Salford systems is capable of automatic variable search, automatic variable transformation and automatic limited interaction search. It outputs variable importance, basis functions and the final model. We used this software to generate the best MARS model and predict blind test data.

4.3 Procedure

MARS software is used here for permeability prediction. The procedure is as follows:

Import data into MARS and setup variables and factors such as interactions, range of variables, max basis functions, maximum interactions etc.

Run the program. Variables are automatically searched and all possible models are calculated. The best model is selected by generalized cross validation (GCV) criteria.

Select the best model. Variable importance, basis functions and final model are output.

Predict the new blind test data set by applying best model to data. MSE and MAE are calculated manually.

4.4 Field Application: Salt Creek Field Unit

After data partitioning using electrofacies characterization, we apply the multivariate adaptive regression method to correlate log-permeability to well logs. We take electrofacies I as an example to illustrate MARS.

Table 4.1 indicates the relative variable importance and Table 4.2 shows the basis functions of electrofacies I. According to Table 4.1 and Table 4.2, we can see that four well logs (RHOB, GR, log10_MSFL, log10_LLD) are selected out of seven logs for data correlation. Log-permeability prediction is a combination of several basis functions and each basis function in this case is a piecewise linear function. Here, we can see that the multivariate adaptive regression splines method is a generalization of piecewise linear regression.

Table 4.1 Relative Variable Importance of Electrofacies I, SCFU

Relative Variable Importance			
5	RHOB	100.000	2.881
1	GR	44.963	2.430
3	LOG10_MSFL	35.831	2.389
2	LOG10_LLD	30.431	2.368
4	DT	0.000	2.316
6	NPHI	0.000	2.316
7	PEF	0.000	2.316

Table 4.2 Basis Functions of Electrofacies I, SCFU

Basis Functions
BF2 = max(0, 2.680 - RHOB);
BF3 = max(0, GR - 4.100);
BF4 = max(0, LOG10_LLD - 1.828);
BF6 = max(0, LOG10_MSFL - 2.033);
BF7 = max(0, 2.033 - LOG10_MSFL);
Y = -2.249 + 24.571 * BF2 - 0.051 * BF3 + 1.533 *B4 + 1.074 * BF6 + 3.423 * BF7;

Repeating the same procedure for all electrofacies; we have the variables selected by MARS out of all electrofacies as shown in Table 4.3. The order of variables gives the relative importance of variables from high to low. Different variable combinations show the dissimilarity of electrofacies. At the same time, we can determine the basis functions of all electrofacies.

Table 4.3 Variable Selection of All Electrofacies, SCFU

Electrofacies	Variables selected (relative importance order)
I	RHOB, GR, log10_MSFL, log10_LLD
II	RHOB, GR, log10_MSFL, log10_LLD
III	RHOB, DT
IV	RHOB, GR, log10_LLD
V	Log10_LLD, NPHI, RHOB
VI	RHOB, GR
VII	Log10_MSFL, log10_LLD, RHOB
VIII	RHOB, log10_LLD, DT, GR, log10_MSFL

Table 4.4 Regression and Prediction Errors of All Electrofacies

	Well	MSE	MAE
Regression error	Training	1.5458	0.9585
Prediction error	G517	2.1267	1.1200
	G520	2.3468	1.1343

Given the basis functions of each electrofacies, we can predict permeability and calculate regression error and prediction error of wells G517 and G520 in Table 4.4. Table 4.5 compares regression and prediction errors of the MARS, ACE and tree methods. As it is shown, MARS yields very good prediction errors compared to the ACE and tree regression methods.

Table 4.5 Comparison of MARS, ACE and Tree Regression

	Error	MARS	ACE	Tree
Regression error	MSE	1.5458	1.58	1.4362
	MAE	0.9585	0.97	0.9454
Prediction error G517	MSE	2.1267	2.25	2.2546
	MAE	1.1200	1.15	1.1798
Prediction error G520	MSE	2.3468	1.74	3.0595
	MAE	1.1343	1.04	1.4078

Figure 4.1 and Figure 4.2 show the predicted permeability compared to measured permeability in two different formats. The correlation coefficient of Figure 4.1 equals

to 0.6951, which indicates a nice prediction by the MARS method.

Figure 4.1 Measured Permeability vs. Predicted Permeability, G517 SCFU, MARS

Figure 4.2 Measured and Predicted Permeability vs. Depth, G517 SCFU, MARS

4.5 Field Application: North Robertson Unit

Here we apply the MARS method to field case NRU. In each electrofacies, variables are selected and basis functions are calculated after generalized cross-validation. Variable importance of all electrofacies is shown in Table 4.6. Table 4.7 summarizes regression error and prediction error of all electrofacies.

Table 4.6 Variables Selected of All Electrofacies, NRU

Electrofacies	Variables selected (relative importance order)
EF1	None
EF2	Log10_LLD, log10_MSFL, GR
EF3	Log10_LLD, NPHI, log10_MSFL
EF4	None
EF5	NPHI, log10_LLD, log10_MSFL, RHOB
EF6	NPHI, DT, log10_LLD
EF7	Log10_LLD, log10_MSFL
EF8	RHOB
EF9	DT, GR

Table 4.7 Regression and Prediction Errors of All Eletrofacies, NRU

	Well	MSE	MAE
Regression error		2.6581	1.2702
Prediction error	W1509	1.4441	0.8401
	W3533	2.8917	1.1732

Figure 4.3 and Figure 4.4 compare the measured permeability and predicted permeability. Just as for stepwise regression and tree regression, MARS is unable to give a good permeability prediction. The correlation coefficient from MARS is 0.1871, compared to stepwise regression (0.1950) and tree regression (0.1339). The poor linear dependence between measured and predicted permeability is shown in Figure 4.3.
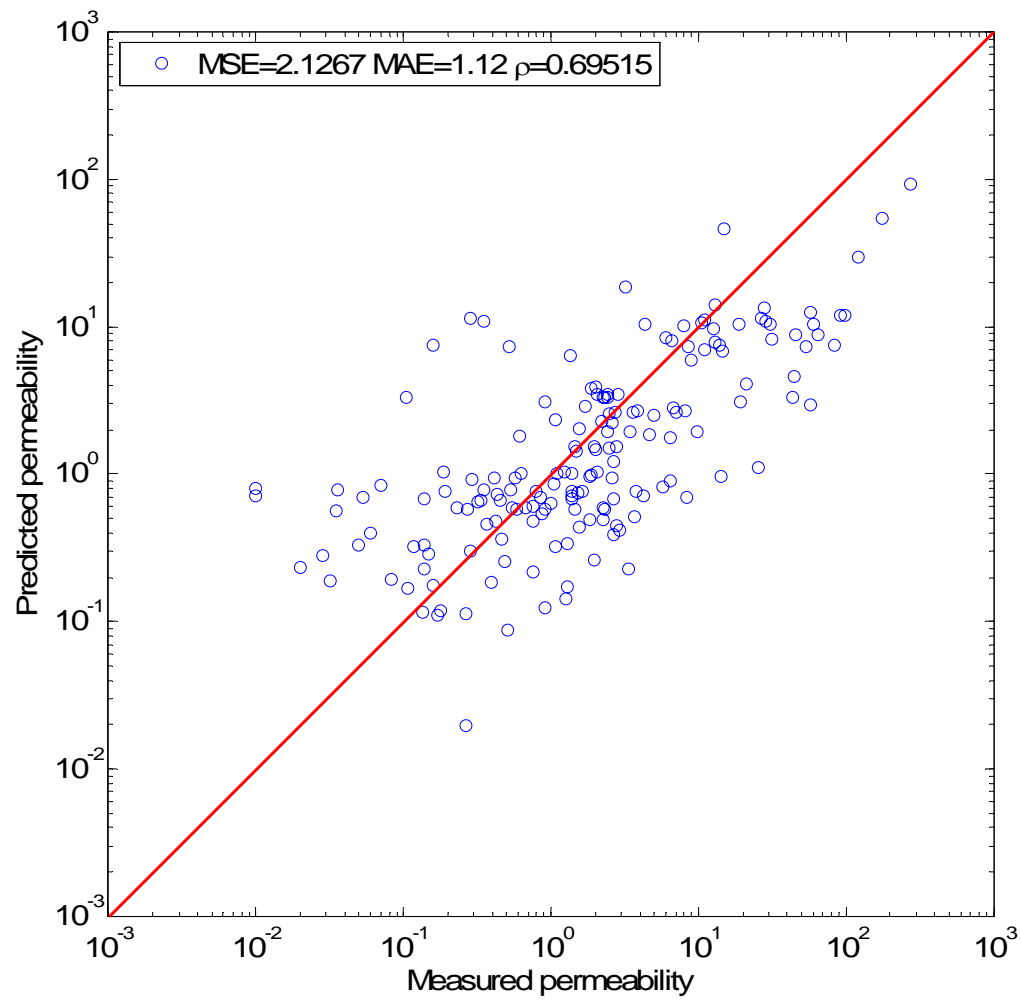

Figure 4.3 Measured Permeability vs. Predicted Permeability, W1509 NRU, MARS

Figure 4.4 Measured and Predicted Permeability vs. Depth, W1509 NRU, MARS

Table 4.8 Comparison of MARS, ACE and Tree Regression, NRU

	Error	MARS	Stepwise	Tree
Regression error	MSE	2.6581	2.8776	2.5210
	MAE	1.2702	1.3225	1.2288
Prediction error W1509	MSE	1.4441	1.3376	1.5286
	MAE	0.8401	0.7661	0.8500
Prediction error W3533	MSE	2.8917	2.8961	2.8074
	MAE	1.1732	1.0985	1.1510

Finally, we compare regression and prediction errors of all three methods: MARS, stepwise regression and tree regression, in Table 4.8. Unexpectedly, stepwise regression outperforms the other two methods in the North Robertson Unit (NRU) case. The reason is that the NRU core permeability data with low accuracy and wide range lead to deviated data partitioning.

4.6 Summary

In this chapter, we discussed the multivariate adaptive regression splines method and apply it to two field cases: Salt Creek Field Unit (SCFU) and North Robertson Unit (NRU). We used the MARS software from Salford Systems to perform this method. The software has built-in functions for the stepwise procedure and generalized cross-validation for knot placement and model reduction. Results from SCFU are summarized in Table 4.5, Figure 4.1 and Figure 4.2. They show better prediction outcomes relative to the ACE algorithm and tree regression.

CHAPTER V

DISCUSSION AND CONCLUSIONS

5.1 Discussion

In this study, we improved permeability estimation based on well logs by focusing on data correlation and variable selection of the permeability prediction process. Variable selection is used to selected appropriate well logs and then data correlation is built between permeability and selected well logs. The following three methods are tested and applied to two highly heterogeneous carbonate reservoirs: Salt Creek Field Unit and North Robertson Unit.

1. Non-parametric regression (ACE) with stepwise procedure and AIC-criterion

2. Tree regression with K-fold cross-validation and pruning

3. The multivariate adaptive regression splines (MARS)

Result from Salt Creek Field Unit shows that permeability estimates is improved by applying variable selection to the permeability prediction process. The correlation coefficients between measured permeability and predicted permeability are close to 0.7, which is an acceptable linear dependency. The mean squared error and the mean absolute error are also showing good permeability estimations. But tree regression is unable to preserve the continuity of permeability and leads to discrete permeability predictions.

In North Robertson Unit, all three methods are unable to capture high permeability. In my opinion, the reason we fail to predict high permeability is not because the multivariate regression methods but the following:

Data partitioning problem;

Low permeability with limited accuracy is hard to predict;

Abnormally high permeability is shown in some low permeability zones. Possible explanation is micro-fractures.

So far, the following conclusions can be drawn based on this work:

1. Variable selection in conjunction with regression methods shows significant potential for permeability predictions in complex carbonate reservoirs.

2. An examination of the error rates for uncored wells in Salt Creek Field Unit indicates that the stepwise with ACE and MARS methods are successful in predicting permeability and tree regression is less successful because it is unable to preserve continuity of permeability.

3. For permeability predictions using a combination of data correlation and variable selection, the success of the method strongly depends on the discriminatory power of the data classification technique. Result of North Robertson Unit shows that the difficulty in indentifying electrofacies in uncored wells can result in the application of incorrect correlation and, consequently, poor permeability predictions.

4. In comparing the relative predictive performance of the three regression methods, the stepwise with ACE method appears to outperform the other two methods.

5.2 Future Work

The idea of variable selection in data correlation is very successful in improving permeability estimation. In statistics, variable selection can also be used in cluster analysis, which is part of our data partitioning. Based on the successful of variable selection in data correlation, I think we can try the variable selection for electrofacies characterization.

Another idea is from MARS software: The software has built-in generalized cross-validation and model selection process. We can try to apply cross-validation to non-parametric regression (ACE). If result is promising, we can implement GRACE with cross-validation for variable selection and model reduction.

REFERENCES

1. Jensen, J. L. and Lake, L. W.: "Optimization of Regression-Based Porosity-Permeability Predictions," *CWLS 10th Symposium, Calgary, Alberta, Canada, September* (1985).

2. Wendt, W. A., Sakurai, S., and Nelson, P. H., *Permeability Prediction from Well Logs Using Multiple Regression*, in *Reservoir Characterization*, L.W. Lake and H.B.J. Caroll (eds), Academic Press, New York (1985).

3. Huang, Z., Shimeld, J., Williamson, M., and Katsube, J.: "Permeability Prediction with Artificial Neural Network Modeling in the Venture Gas Field, Offshore Eastern Canada," *Geophysics* (1996), **61**, No.2, 422-436.

4. Barman, I., Sharma, A., Walker, R., and Datta-Gupta, A.: "Permeability Predictions in Carbonate Reservoirs Using Optimal Non-parametric Transforms: An Application at the Salt Creek Field, Kent County, TX," *SPE39667 paper presented at the 1998 SPE/DOE Symposium on Improved Oil Recovery, Tulsa, OK, April 19-22* (1998).

5. Datta-Gupta, A., Xue, G., and Lee, S. H., *Nonparametric Transformations for Data Correlation and Integration: From Theory to Practice*, in *Reservoir Characterization: Recent Advances*, AAPG Datapages (1999).

6. Xue, G., Datta-Gupta, A., Valko, P., and Blasingame, T.: "Optimal Transformations for Multiple Regression: Application to Permeability Estimation from Well Logs," *SPE Reservoir Evaluation & Engineering* (1997), **12**, No.2, 85-94.

7. Davies, D. K. and Vessell, R. K.: "Flow Unit Characterization of a Shallow Shelf Carbonate Reservoir: North Robertson Unit, West Texas," *SPE35433*

paper presented at the 1996 SPE/DOE Symposium on Improved Oil Recovery, Tulsa, OK, April 21-24 (1996).

8.   Moline, R. G. and Bahr, J. M.: "Estimating Spatial Distributions of Heterogeneous Subsurface Characteristics by Regionalized Classification of Electrofacies," *Mathematical Geology* (1995), **27**, No.1, 3-22.

9.   Hastie, T. and Tibshirani, R., Generalized Additive Models. Chapman & Hall/CRC, New York (1990).

10.   Lee, S. H., Arun, K., and Datta-Gupta, A.: "Electrofacies Characterization and Permeability Predictions in Complex Reservoirs," *SPE Reservoir Evaluation & Engineering* (2002), **5**, No.3, 237-248.

11.   Mathisen, T., Lee, S. H., and Datta-Gupta, A.: "Improved Permeability Estimates in Carbonate Reservoirs Using Electrofacies Characterization: A Case Study of the North Robertson Unit, West Texas," *SPE Reservoir Evaluation & Engineering* (2003), **6**, No.3, 176-184.

12.   Wang, D. and Murphy, M.: "Estimating Optimal Transformations for Multiple Regression Using the ACE Algorithm," *Journal of Data Science* (2004), **2**, No.4, 329-346.

13.   Maindonald, J. H. and Braun, J., Data Analysis and Graphics Using R: An Example-based Approach. Cambridge University Press, New York (2007).

14.   Venables, W. N. and Ripley, B. D., Modern Applied Statistics with S-Plus. Springer, Berlin (2002).

15.   Friedman, J., Barron, A., Xiao, X., Golubev, G., Hasminskii, R., Breiman, L., Bujaz, A., Duffy, D., Hastie, T., and Tischirani, R.: "Multivariate Adaptive Regression Splines: Discussions," *Annals of Statistics* (1991), **19**, No.1, 1-141.

16.   Friedman, J. H.: "Rejoinder: Multivariate Adaptive Regression Splines," *Annual of Statistics* (1991), **19**, No.1, 123-141.

17. Friedman, J. and Roosen, C.: "An Introduction to Multivariate Adaptive Regression Splines," *Statistical Methods in Medical Research* (1995), **4**, No.3, 197-217.

18. Dorfman, M. H., Newey, J. J., and Coates, G. R., *New Techniques in Lithofacies Determination and Permeability Prediction in Carbonates using Well Logs*, in *Geological Applications of Wireline Logs*, A. Hurst, Lovell, M. A., and Motron A. C. (eds) Geological Society London Special Publications (1990). P. 113.

19. Abbasadeh, M., Fujii, H., and Fujimoto, F.: "Permeability Prediction by Hydraulic Flow Units: Theory and Applications," *SPE Formation Evaluation* (1996), **11**, No.4, 263-271.

20. Amaefule, J., Altunbay, M., Tiab, D., Kersey, D., and Keelan, D.: "Enhanced Reservoir Description: Using Core and Log Data to Identify Hydraulic (Flow) Units and Predict Permeability in Uncored Intervals/Wells," *SPE26436 paper presented at the 68th SPE Annual Technical Conference and Exhibition (1993), Houston, Texas, Oct. 3-6.* (1993).

21. Perez, H., Datta-Gupta, A., and Mishra, S.: "The Role of Electrofacies, Lithofacies, and Hydraulic Flow Units in Permeability Prediction From Well Logs: A Comparative Analysis Using Classification Trees," *SPE Reservoir Evaluation & Engineering* (2005), **8**, No.2, 143-155.

22. Bucheb, J. and Evans, H.: "Some Applications of Methods Used in Electrofacies Identification," *LOG ANALYST* (1994), **35**, No.6, 14-14.

23. Kalkomey, T. C.: "Potential Risks When Using Seismic Attributes as Predictors of Reservoir Properties," *The Leading Edge* (1997), **16**, No.3, 247-252.

24. Christensen, R., Analysis of Variance, Design and Regression: Applied Statistical Methods. Chapman & Hall, New York (1996).

25. Lee, S. H., Arun, K., and Datta-Gupta, A.: "Electrofacies Characterization and

Permeability Predictions in Complex Reservoirs," *SPE Reservoir Evaluation & Engineering* (2002). **5**, No.3, 237-248

26. Breiman, L., Friedman, J., Pregibon, D., Vardi, Y., Buja, A., Kass, R., Fowlkes, E., and Kettenring, J.: "Estimating Optimal Transformations for Multiple Regression and Correlation. Comments and Rejoinder," *Journal of the American Statistical Association* (1985), **80**, No.391, 580-619.

VITA

Jiang Xie received his B.S. degree in Polymer Science & Engineering from University of Science and Technology of China in May 2005, and his M.S. degree in Petroleum Engineering from Texas A&M University in December 2008.

His current address: Texas A&M University

                        Harold Vance Department of Petroleum Engineering

                        3116 TAMU

                        College Station, TX 77843-3116

                        U.S.A.

Email: jiang.xie@pe.tamu.edu