SPE/DOE 35412

Optimal Transformations for Multiple Regression: Application to Permeability Estimation from Well Logs

Guoping Xue, Akhil Datta-Gupta, Peter Valko, and Tom Blasingame, Texas A&M University.

SPE Members

This paper was prepared for presentation at the Improved Oil Recovery Symposium held in Tulsa, Oklahoma, 21 April 1996

Abstract

Conventional multiple regression for permeability estimation from well logs requires a functional relationship to be presumed. Due to the inexact nature of the relationship between petrophysical variables, it is not always possible to identify the underlying functional form between dependent and independent variables in advance. When large variations in petrological properties are exhibited, parametric regression often fails or leads to unstable and erroneous results, especially for multivariate cases.

In this paper we describe a nonparametric approach for estimating optimal transformations of petrophysical data to obtain the maximum correlation between observed variables. The approach does not require *a priori* assumptions of a functional form and the optimal transformations are derived solely based on the data set. An iterative procedure involving the *alternating conditional expectation* (ACE) forms the basis of our approach. The power of ACE is illustrated using synthetic as well as field examples. The results clearly demonstrate improved permeability estimation by ACE compared to conventional parametric regression methods.

Introduction

A critical aspect of reservoir description involves estimating permeability in uncored wells based on well logs and other known petrophysical attributes. A common approach is to develop a permeability-porosity relationship by regressing on data from cored wells and then, to predict permeability in uncored wells from well logs.[1,2] Multiple regression is used when large variations in petrological properties exist (e. g. a wide range in grain sizes, high degree of cementation, diagenetic alteration etc.) and a simple permeability-porosity relationship no longer holds good. However, there are several limitations to such an approach. Many of these arise from the inexact nature of the relationship between petrophysical variables and *a priori* assumptions regarding functional forms used to model the data -- all leading to biased estimates. When prediction of permeability extremes is a major concern, the high and low values are enhanced through a weighting scheme in the regression. Besides being subjective in nature, such weighting can cause the prediction to become unstable which leads to erroneous results. Most importantly, conventional regression assumes independent variables to be free of error, which is highly optimistic for geologic and petrophysical data.

Jensen and Lake[2] introduced power transformations for optimization of regression-based permeability-porosity predictions. The underlying theory is that if the joint probability distribution function (j.p.d.f.) of two variables is binormal, the relationship will be linear.[3] Several methods exist to estimate the exponents for power transformation. One method, described by Emerson and Stoto[4] and adopted by Jensen and Lake,[2] is based on symmetrizing the probability distribution function (p.d.f.). Another method is a trial-and-error approach based on a normal probability plot of the data. By power transforming permeability and porosity separately, the authors are able to improve permeability-porosity correlations. However, using a trial-and-error method for selecting exponents for power transformation is time consuming, and symmetrizing the p.d.f. does not necessarily guarantee a binormal distribution of transformed variables. In addition, there are no indications as to whether power transformations will work for multivariate cases.

References and illustrations at end of paper.

Nonparametric regression techniques[5] based on statistical and optimization theory have been developed to offer much more flexible data analysis tools for exploring the underlying relationships between dependent and independent variables. In this paper, we describe a very general and computationally-efficient nonparametric regression algorithm called ACE.[6] The algorithm provides a method for estimating transformations in multiple regression without prior assumptions of a functional relationship. The ACE transformations are shown to be optimal and yield maximum correlation between the variables in the transformed space. Unlike conventional multiple regression, the proposed approach allows for data correction/equilibration for the dependent as well as independent variables. The power of the method lies in its ease of use, particularly for multivariate regression, and its ability to identify and correct for outliers without subjective assumptions.

The organization of this paper is as follows. First, we discuss the underlying theory of the ACE method and its implementation. Second, we apply ACE to synthetic examples to illustrate function identification during multiple regression and to also show data correction features. Finally, we present field examples involving petrophysical data from three formations -- the Spraberry formation in Texas (a highly interbedded turbidite sandstone), the Admire sand in the El Dorado Field in Kansas (a shallow delta sand), and the Schneider Buda Field in Texas (a Cretaceous reefal limestone). In the Appendix we describe a data smoothing technique called Supersmoother[8] which is used to replace conditional expectation calculations for finite data set.

ACE Technique

The ACE algorithm, originally proposed by Breiman and Friedman,[6] provides a method for estimating optimal transformations for multiple regression that result in a maximum correlation between a dependent (response) random variable and multiple independent (predictor) random variables. A brief description of the theory of ACE and its implementation as applied to continuous random variables are given in the following sections. See Breiman and Friedman[6] for further details.

Optimal Transformations. Optimal transformations minimize the variance of a linear relationship between the transformed dependent and independent variables. For a given set of a dependent random variable Y and independent random variables $X_1,...,X_p$, we first define arbitrary measurable mean-zero transformations $\theta(Y), \phi_1(X_1),...,\phi_p(X_p)$. The error (e^2) not explained by a regression of the transformed dependent variable on the sum of transformed independent variables is (under the constraint, $E[\theta^2(Y)] = 1$)

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

Transformations $\theta^*(Y), \phi_1^*(X_1),...,\phi_p^*(X_p)$ are said to be optimal for regression if they satisfy the following

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

We also define a correlation coefficient between the transformed dependent variable and the sum of transformed independent variables (with the constraints, $E[\theta^2(Y)] = 1$ and $E[\phi_s{}^2(X)] = 1$) as follows

$$\rho(\theta,\phi_s) = E[\theta(Y)\phi_s(X)] \qquad (3)$$

where $\phi_s(X) = \sum_{l=1}^{p} \phi_l(X_l)$.

Transformations $\theta^{**}(Y), \phi_1^{**}(X_1),...,\phi_p^{**}(X_p)$ are said to be optimal for correlation if

$$\rho^*(\theta^{**},\phi_s^{**}) = \max_{\theta,\phi_1,...,\phi_p} \rho(\theta,\phi_s) \qquad (4)$$

It can be shown[6] that if $\theta^{**}(Y), \phi_1^{**}(X_1),...,\phi_p^{**}(X_p)$ are optimal for correlation, then $\theta^*(Y) = \theta^{**}(Y)$, $\phi_1^*(X_1) = \rho^*\phi_1^{**}(X_1)$, ..., $\phi_p^*(X_p) = \rho^*\phi_p^{**}(X_p)$ are optimal for regression, and vice versa. The minimum regression error and maximum correlation coefficient are related by $e^{*2} = 1 - \rho^{*2}$.

Proof of the existence of optimal transformations can be found in the paper by Breiman and Friedman[6].

The ACE Algorithm. For clarity and without loss of generality, let us consider a bivariate case $(p = 1)$. Our objective is to minimize the following equation (with the constraint, $E[\theta^2(Y)] = 1$)

$$e^2(\theta,\phi) = E\{[\theta(Y) - \phi(X)]^2\} \qquad (5)$$

The minimization of e^2 with respect to $\phi(X)$ for a given $\theta(Y)$ yields

$$\phi(X) = E[\theta(Y)|X] \quad (6)$$

Similarly, the solution to the minimization of e^2 with respect to $\theta(Y)$ for a given $\phi(X)$ is

$$\theta(Y) = E[\phi(X)|Y]/\|E[\phi(X)|Y]\| \quad (7)$$

Eqs. 6 and 7 form the basis of the ACE algorithm. The procedure involves an iterative calculation of conditional expectations as follows

Step 1: set $\theta^0(Y) = Y/\|Y\|$, as initial value, and calculate

$\phi^0(X) = E[\theta^0(Y)|X]$

Then, for $k \geq 1$

Step 2: calculate $\phi^k(X) = E[\theta^{k-1}(Y)|X]$

Step 3: calculate $\theta^k(Y) = E[\phi^k(X)|Y]/\|E[\phi^k(X)|Y]\|$

Step 4: calculate $\Delta e^2 = |e^2(\theta^k,\phi^k) - e^2(\theta^{k-1},\phi^{k-1})|$

Stop if $\Delta e^2 \leq \delta$ (a given tolerance), otherwise repeat Step 2 through Step 4.

The final $\theta^k(Y)$ and $\phi^k(X)$ are the solutions to the optimal transformations $\theta^*(Y)$ and $\phi^*(X)$.

The discussion given above pertains to continuous random variables with known conditional expectations. More generally, we will be given a finite data set with n observations $\{y_i, x_{1i}, ..., x_{pi}, 1 \leq i \leq n\}$ which are assumed to be sampled from random variables $Y, X_1, ..., X_p$. Since the j.p.d.f. for a finite data set is rarely known, calculation of conditional expectations in the ACE algorithm is replaced by data smooths.[6] Thus, the transformations derived by ACE, $\{\theta^*(y_i), \phi_1^*(x_{1i}), ..., \phi_p^*(x_{pi}), 1 \leq i \leq n\}$ for a finite data set are the estimates of optimal transformations and are in discrete form.

We include in the Appendix a short description of a data smoothing algorithm called Supersmoother that is used in lieu of conditional expectations for finite data sets.

ACE for Estimation. The following equation is used to estimate or predict dependent variable, y_j^{pre} for any given data point $\{x_{1j}, ..., x_{pj}\}$ involving p - independent variables

$$y_j^{pre} = \theta^{*-1}[\sum_{l=1}^{p} \phi_l^*(x_{lj})] \quad (8)$$

The calculation involves p forward transformations of $\{x_{1j}, ..., x_{pj}\}$ to $\{\phi_1^*(x_{1j}), ..., \phi_p^*(x_{pj})\}$, and a backward transformation, $\theta^{*-1}[\sum_{l=1}^{p} \phi_l^*(x_{lj})]$.

Applications

In this section we apply the ACE technique to synthetic and field examples. Synthetic examples illustrate the utility of ACE for function identification and data correction. Field examples are used to test ACE for permeability estimation from porosity and well logs.

Synthetic Examples. We discuss here two examples to illustrate how the ACE algorithm can be used to identify the functional relationship between dependent and independent variables. A data correction technique is proposed in conjunction with the ACE algorithm when both dependent and independent variables are subject to errors. We describe the procedure for the data correction and give an example to illustrate the use of the technique.

Function Identification. Our first example is a bivariate case similar to a case given by Breiman and Friedman[6]. A data set with 200 observations are simulated from the following model

$$y_i = \exp[\sin(2\pi x_i) + \varepsilon_i / 2] \qquad (1 \leq i \leq 200) \quad (9)$$

where x_i is drawn from a uniform distribution $U(0,1)$ and ε_i is independently drawn from a standard normal distribution $N(0,1)$. **Fig. 1a** is a scatterplot of y_i versus x_i. The plot itself does not reveal a functional relationship between the dependent and independent data observations. In this situation the direct use of parametric regression is impossible.

Rearranging Eq. 9 as

$$\ln(y_i) = \sin(2\pi x_i) + \varepsilon_i / 2 \ , \quad (10)$$

we are able to identify that the optimal transformations of dependent and independent variables will have the following forms

$$\theta^*(y_i) = \ln(y_i)$$
$$\phi^*(x_i) = \sin(2\pi x_i) \quad (11)$$

Substituting Eq. 11 into Eq. 10 results in

$$\theta^*(y_i) = \phi^*(x_i) + \varepsilon_i / 2 \quad (12)$$

To demonstrate that the ACE algorithm can estimate the above optimal transformations, we applied the algorithm to our synthetic data set. **Figs. 1b** and **1c** show the optimal transformations of y_i and x_i derived by ACE. Clearly, ACE is able to identify logarithmic function as the optimal transformation of the dependent variable and sine function as the optimal transformation of the independent variable. **Fig. 1d** shows a plot of $\theta^*(y_i)$ versus $\phi^*(x_i)$. A linear regression on the transformed data yields the following

$$\theta^*(y_i) \approx 1.093\,\phi^*(x_i) \tag{13}$$

which is a very close estimate of $\theta^*(y_i) = \phi^*(x_i)$, indicating that the transformations are, indeed, optimal. We note that the original y_i and x_i are preprocessed to satisfy a zero mean for both y_i and x_i and a unit variance for y_i prior to the ACE transformations. Hence, the optimal transformations estimated by ACE, as shown in **Figs. 1b and 1c,** do not have the exact numerical relations as in Eq. 11.

Our second example is a case with multiple independent variables. We believe that the real power of ACE lies in such multivariate cases where a conventional regression is often inadequate, and can lead to erroneous results. Our example involves 300 observations generated using the following model

$$y_i = x_{1i} + x_{2i}^2 + x_{3i}^3 + 0.1\varepsilon_i \qquad (1 \le i \le 300) \tag{14}$$

where x_{1i}, x_{2i}, and x_{3i} are independently drawn from a uniform distribution $U(-0.5,0.5)$, and ε_i is drawn from a standard normal distribution $N(0,1)$. **Fig. 2a through Fig. 2c** show plots of y_i versus x_{1i}, x_{2i}, and x_{3i}, respectively. Except for y_i versus x_{1i}, the functional relationships between the dependent variable y_i and independent variables x_{2i} and x_{3i} can not be identified from the scatterplots.

The optimal transformations for y_i, x_{1i}, x_{2i} and x_{3i} derived using ACE are plotted in **Fig. 2d through Fig. 2g**. As expected, the transformations for both y_i and x_{1i} yield essentially straight lines. The transformation for x_{2i} reveals a quadratic function and the transformation for x_{3i} reveals a cubic function. Thus ACE is able to identify the following optimal transformations

$$\theta^*(y_i) = y_i$$
$$\phi_1^*(x_{1i}) = x_{1i}, \quad \phi_2^*(x_{2i}) = x_{2i}^2, \quad \phi_3^*(x_{3i}) = x_{3i}^3 \tag{15}$$

This is, indeed, remarkable considering that the individual scatterplots hardly reveal any such relationships.

A plot of transformed y_i versus the sum of transformed x_{1i}, x_{2i} and x_{3i} is shown in **Fig. 2h**. The relationship can be fitted approximately by

$$\theta^*(y_i) = \phi_1^*(x_{1i}) + \phi_2^*(x_{2i}) + \phi_3^*(x_{3i}) \tag{16}$$

which is exactly optimal.

These two examples clearly demonstrate the power of the ACE algorithm for function identification through optimal transformations.

Data Correction. The purpose of data correction is to remove any error or noise introduced during data measurements. For a noise-free and perfectly correlated data set $\{y_i, x_{1i}, \ldots, x_{pi},\ 1 \le i \le n\}$, we expect that the ACE derived optimal transformations will satisfy exactly the following

$$\theta^*(y_i) = \sum_{l=1}^{p} \phi_l^*(x_{li}) \tag{17}$$

Hence, in the transformed space, a plot of $\theta^*(y_i)$ versus $\sum_{l=1}^{p} \phi_l^*(x_{li})$ is a straight line with unit slope, except for the contaminated data which will deviate from the straight line such as point A in **Fig. 3**. Our procedure for data correction is to move the deviated point to the unit slope line by adding a correction. In addition, we require that the movement distance be minimum. The solution to this geometric problem is

$$\Delta\theta^*(y_i) = -\frac{1}{2}\Delta E_i$$
$$\Delta\phi_l^*(x_{li}) = \frac{1}{2p}\Delta E_i \qquad (1 \le l \le p) \tag{18}$$

where

$$\Delta E_i = \theta^*(y_i) - \sum_{l=1}^{p} \phi_l^*(x_{li}) \tag{19}$$

Fig. 3 is a schematic drawing that shows the data correction procedure. In the transformed space, the data pair located at point A is corrected to point A′ according to Eqs. 18 and 19. As such, this procedure allows for correction of dependent as well as independent variables. By comparison, the conventional methods that associate all errors to the dependent

variable will correct point A to point A''. Finally, we determine the corresponding data correction in the original space by back transformation as follows

$$y_i^{corr} = y_i + \Delta y_i = \theta^{*-1}[\theta^*(y_i) + \Delta\theta^*(y_i)]$$
$$x_{li}^{corr} = x_{li} + \Delta x_{li} = \phi_l^{*-1}[\phi_l^*(x_{li}) + \Delta\phi_l^*(x_{li})] \quad (1 \le l \le p) \tag{20}$$

where y_i^{corr} and x_{li}^{corr}, $1 \le l \le p$ are corrected values.

An illustration of data correction in two-dimension is shown in **Figs. 4a and 4b**. **Fig. 4a** shows the noisy data generated by adding random gaussian noise to the underlying plane. **Fig. 4b** shows the location of the data points after correction using ACE. For all practical purposes, ACE is able to bring the data points back to the plane.

Field Examples. We have applied the ACE technique to permeability estimation for three field examples. In the first two examples ACE is used to derive permeability-porosity correlations. In third example we discuss permeability estimation directly from well logs using ACE.

Spraberry Formation in Parks Field. Spraberry sandstones in the Parks Field, West Texas, are channel turbidites deposited in a submarine fan environment. The data used for this study consist of 402 core plug porosities and permeabilities. The mean porosity and permeability values for this data set are 7.7% and 0.13 md, respectively. A detailed description of the reservoir and its depositional environment can be found in Yale[7].

Fig. 5a is a scatterplot of permeability versus porosity. Our goal is to find the transformations that will result in a maximum correlation between core permeability and porosity. **Figs. 5b and 5c** show the optimal transformations of permeability and porosity derived using ACE. The shape of these transformations clearly indicates that the conventional $\log(k) \sim \phi$ regression model is inadequate for this data set. **Fig. 5d** shows a plot of transformed permeability versus transformed porosity. A linear regression in the transformed space gives a correlation $R^2 = 0.799$. **Fig. 5e** compares the original permeability with the corrected permeability after data correction. Note that the solid line in **Fig. 5e** corresponds to a unit slope straight line in the transformed space.

For comparison, we performed a regression based on a logarithmic transformation of the permeability, $\log(k) = a_1\phi + a_2$. **Fig. 5f** is a scatterplot of $\log(k)$ versus ϕ, with a linear correlation $R^2 = 0.692$, which is much lower than that obtained by ACE.

Another method commonly used in data analysis is normal score transformation. The porosity and permeability data are transformed to their normal scores and plotted in **Fig. 5g**. A linear regression of transformed permeability on transformed porosity results in a correlation $R^2 = 0.658$, which is close to that of logarithmic transformation. The results from these three transformation methods are summarized in **Table 1**. ACE is clearly the preferred choice and results in a significant improvement in correlation compared to conventional methods.

Admire Sand in El Dorado Field. The second field example is taken from the Admire sand in the El Dorado Field, Kansas. The Admire sand was deposited in a shallow delta environment. The available data for this study consist of 162 core plug permeabilities and porosities. A detailed description of the data can be found in Jensen and Lake.[2]

Fig. 6a is a scatterplot of permeability versus porosity. A conventional regression of the data yields the following relationship (**Fig. 6b**)

$$\log(k) = 0.217\phi - 3.393 \tag{21}$$

with $R^2 = 0.626$. From **Fig. 6b**, it appears that the variability of permeability is much higher at low porosities compared to high porosities. The low correlation R^2 value reflects the high variability of permeability at lower porosities.

We next performed a normal score transformation of the data resulting in the following $k - \phi$ relationship (**Fig. 6c**)

$$NS(k) = 0.862\,NS(\phi) \tag{22}$$

with $R^2 = 0.743$.

Power transformations using the Jensen and Lake[2] method give the following relation (**Fig. 6d**)

$$k^{0.61} = 0.0864 \times 10^{-5}\phi^{5.3} + 3.889 \tag{23}$$

with $R^2 = 0.758$.

Finally, we applied ACE to the same permeability and porosity data set and obtained the following relationship of the transformed data (**Fig. 6e**)

$$\theta^*(k) = 1.048\,\phi^*(\phi) \tag{24}$$

with $R^2 = 0.778$.

The close performance of ACE and power transformation method for this case can be explained by examining the transformations for permeability and porosity derived by ACE as shown in **Figs. 6f and 6g**. The transformation for permeability can be fitted by the following power function

$$\theta^*(k) = 0.154k^{0.45} - 2.150\ , \tag{25}$$

and the transformation for porosity can be fitted by a power function as follows

$$\phi^*(\phi) = 0.761 \times 10^{-5}\phi^{3.8} - 2.037 \qquad (26)$$

The exponents in Eqs. 25 and 26 are different from those calculated using the approach suggested by Jensen and Lake[2] and result in a correlation $R^2 = 0.772$. A comparison of correlation statistics from all the transformation methods discussed above is given in **Table 2.** The results show that the transformations derived by ACE are, indeed, optimal. Furthermore, ACE provides a direct method for estimating the exponents of power transformation as opposed to conventional trial-and-error approach.

Schneider Buda Field. Our third field example is from a Cretaceous reefal limestone in Schneider Buda Field, Wood County, Texas. The data set includes 104 core plug porosity and permeability measurements and well log data from the same well. Our goal is to apply ACE for permeability estimation directly from well logs.

We selected four well log variables: sonic travel time (Δt), density derived porosity (ϕ_d), gamma ray (GR), and resistivity ratio (RR) which is the ratio of deep resistivity reading (LLD) to shallow resistivity reading ($MSFL$). **Fig. 7a through Fig. 7d** show scatterplots of logarithm of core permeability versus Δt, ϕ_d, GR, and RR, respectively. A linear regression of $\log(k)$ on individual well log variables results in a maximum correlation $R^2 = 0.477$ with Δt and a minimum correlation $R^2 = 0.0547$ with GR. Such low correlation of permeability with individual well logs are not unusual for a reefal limestone with complicated lithology.

The optimal transformations for permeability and four selected well log variables are shown in **Fig. 7e through Fig. 7i.** The irregular shapes of the estimated transformations further reveal the complicated relationship between core permeability and well logs for this particular formation. The magnitude of the transformed travel time $\phi_1^*(\Delta t)$ is the highest compared to other transformed well log variables. It indicates that travel time has the highest correlation with core permeability which is consistent with the results from individual correlation between logarithmic permeability and well log variables.

Fig. 7j shows transformed core permeability versus the sum of transformed well log variables. A linear regression in the transformed space results in

$$\theta^*(k) = 0.979\,[\phi_1^*(\Delta t) + \phi_2^*(\phi_d) + \phi_3^*(GR) + \phi_4^*(RR)] \qquad (27)$$

with $R^2 = 0.619$, which is much higher than the ones obtained by correlating core permeability with individual well log variables.

For comparison purposes, a conventional linear regression model was used to correlate permeability with the well log variables resulting in the following relationship

$$\log(k) = 0.151\Delta t - 0.0196\phi_d - 0.0392\,GR + 0.0222\,RR - 7.729 \qquad (28)$$

with $R^2 = 0.494$. Clearly, ACE outperforms conventional regression by a wide margin.

Finally, we predicted permeability from the well log data using the following equation derived by ACE

$$k_j^{pre} = \theta^{*-1}[\phi_1^*(\Delta t_j) + \phi_2^*(\phi_{dj}) + \phi_3^*(GR_j) + \phi_4^*(RR_j)] \qquad (29)$$

where k_j^{pre} is the predicted permeability for a given well log reading $\{\Delta t_j, \phi_{dj}, GR_j, RR_j\}$.

Fig. 7k shows a comparison of core permeability with predicted permeability from well logs. Considering the scatter in the original data, the prediction is quite satisfactory.

Discussion

In this paper we have introduced a nonparametric approach, called ACE, as a tool for data analysis and building correlation. The power of ACE has been demonstrated through application to synthetic and field examples. In particular, for cases involving multiple independent variables ACE provides a powerful and easy to use technique for building correlations without the problems associated with conventional multiple regression. The transformations derived by ACE can also facilitate function identification during multiple regression. Our results show that ACE can improve permeability estimation under different depositional environment.

Application of ACE can be extended to many other areas where multiple regression and correlation technique are needed. Of particular interest is reservoir characterization using geostatistical approach. To incorporate various hard (such as core data) and soft data (such as 3-D seismic data) into reservoir characterization, cokriging and cosimulation techniques are often used. The application of the ACE transformations to the hard and soft data should facilitate the computation involved in the use of cokriging and cosimulation method and improve reservoir characterization.

Besides all the strength of ACE as a data analysis tool some important issues related to its application must be

pointed out. Although ACE provides a fully automated approach to estimating optimal transformations, it is not meant to exclude the heuristic reasoning based on the experience gained in the data analysis. In fact, ACE has the option to incorporate a presumed functional form in the model if necessary. A combination of parametric with nonparametric (may be called semi-parametric) regression may be more desirable under certain circumstances. It should be emphasized that the success of the ACE algorithm, like any other regression methods, relies on the quality of the data and underlying association between the dependent and independent variable. The importance of good quality data can not be overemphasized.

Conclusions

1. ACE is a powerful, versatile, and easy-to-use tool for building correlations for petrological variables.

2. ACE provides a fully automated method for estimating optimal transformations for multiple regression. The transformations result in a maximum correlation between observed data.

3. The transformations derived by ACE provide a direct approach to identifying functional relationships between dependent and independent variables. This can be of immense use during conventional parametric regression, especially for cases involving multiple independent variables.

4. The data correction technique incorporated in ACE allows correcting noises for dependent and independent variables.

5. We have demonstrated using field examples that ACE performs significantly better than conventional logarithmic, normal score, and power transformations for estimating permeability from porosity.

6. The power of ACE lies in its ability to directly incorporate various well logs into permeability estimation. This allows for large variations in petrological properties. The application of ACE to the Schneider Buda Field improved correlation of permeability with well logs to $R^2 = 0.619$, compared to $R^2 = 0.494$ by a conventional linear regression method.

Nomenclature

$\hat{l}_{cv}^2$ = average squared residual
e^2 = ACE regression error
ESE = expected squared error
GR = gamma ray, API units
j.p.d.f. = joint probability density function
k = permeability, md
p.d.f. = probability density function
R^2 = correlation R^2 value
RR = resistivity ratio
$X, X_1, \ldots, X_p$ = independent or predictor random variables
$x_1, \ldots, x_p$ = independent data observations
y = dependent data observations
Y = dependent or response random variable
δ = error tolerance
ΔE = data noise or error
Δt = sonic travel time, μs/ft
ε = Gaussian random noise
ϕ = porosity, percent
ϕ_d = density derived porosity, percent
ρ = correlation coefficient
σ^2 = variance

operators

$E(\bullet)$ = mathematical expectation
$E(\bullet|\bullet)$ = conditional mathematical expectation
$\|\bullet\|$ = norm operation
$\theta(\bullet)$ = transformation for independent variable
$\theta^*(\bullet)$ = optimal transformation for independent variable
$\phi_l(\bullet)$ = transformation for independent variable l
$\phi_l^*(\bullet)$ = optimal transformation for independent variable l
$s(\bullet)$ = smoothing

subscripts and superscripts

i,j = indices for data observations
l = index for independent or predictor variables
corr = corrected
pre = predicted
s = summation
cv = cross-validation

Acknowledgments

Akhil Datta-Gupta would like to thank Dr. D.W. Vasco at the Lawrence Berkeley National Laboratory for introducing him to the ACE technique. The authors thank Dr. Larry W. Lake of UT Austin for providing the El Dorado data set and Dr. Richard Strickland of Cawley, Gillespie & Associates for providing the Schneider Buda data.

References

1. Wendt, W.A., Sakurai, S., and Nelson, P.H.: "Permeability Prediction From Well Logs Using Multiple Regression," *Reservoir Characterization,* Edited by Lake, L.W. and Carroll, H.B. Jr., Academic Press, Inc. Orlando, Florida (1986) 659.
2. Jensen, L.J. and Lake, L.W.: "Optimization of Regression-Based Porosity-Permeability Predictions," CWLS 10th Symposium, Calgary, Alberta, Canada, Sep. 29-Oct. 2, 1985.
3. Hald, A.: *Statistical Theory with Engineering Applications*, John Wiley and Sons, New York (1952).

4. Emerson, J. and Stoto, M.: "Exploratory Methods for Choosing Power Transformations," Journal of the American Statistical Association (March, 1982) 103.
5. Hardle, W.: *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, England (1990).
6. Breiman, L. and Friedman, J.H.: "Estimating Optimal Transformations for Multiple Regression and Correlation," Journal of the American Statistical Association (September, 1985) 580.
7. Yale, M.W.: "Depositional Environment and Reservoir Morphology of Spraberry Sandstones, Parks Field, Midland County, Texas," M.S. Thesis, Texas A&M University, College Station, Texas (1986).
8. Friedman, J.H. and Stuetzle, W.: "Smoothing of Scatterplots," Technical Report ORION006, Dept. of Statistics, Stanford University, California (July, 1982).

Appendix-Supersmoother[8]

Concept of data smoother. As we mentioned in the text, when the ACE algorithm is implemented on a finite data set, a data smoother instead of conditional expectation is used. Friedman and Stuetzle[8] define a smoother as a procedure that operates on bivariate data set $\{y_i, x_i,\ 1 \le i \le n\}$ and produces a decomposition

$$y_i = s(x_i) + r_i \quad (1 \le i \le n) \tag{A1}$$

where s is a smooth function or simply smooth, and r_i are residuals. If one assumes that the data y_i is generated from a smooth function plus random noise

$$y_i = f(x_i) + \varepsilon_i, \tag{A2}$$

then smoothers can be viewed as curve estimators and smooth s is considered as an estimate for f. Examples of commonly used data smoothers are local averaging, kernel, etc. A smoother with desirable properties called a Supersmoother is used in Breiman and Friedman's ACE transformations. The following is a brief description of the Supersmoother developed by Friedman and Stuetzle[8].

Supersmoother. We assume that the data are generated according to Eq. A2. The goal of Supersmoother is to find a procedure that can approximate function f as closely as desired given a dense enough data set and without any condition on f apart from being smooth. Supersmoother starts from a simple moving average in which the smoothed value $s(x_i)$ at point x_i is given as

$$s(x_i|k) = \frac{1}{k}\sum_{j=i-k/2}^{i+k/2} y_j \tag{A3}$$

where k is chosen as an odd number and is called span. At the boundaries where it is not possible to keep the data points symmetric around smoothing point only available data points are used for averaging.

The most important parameter that needs to be determined for a local averaging smoother is span. The span controls the trade off between bias and variance of the estimate. If the errors ε_i in Eq. A2 are independent and identically distributed with zero mean and variance σ^2, then expected squared error at point x_i with a given span k can be expressed as[8]

$$ESE(x_i|k) = [f(x_i) - \frac{1}{k}\sum_{j=i-k/2}^{i+k/2} f(x_j)]^2 + \frac{1}{k}\sigma^2 \tag{A4}$$

The first term in the right hand of Eq. A4 can be expanded using a Taylor series as

$$[f(x_i) - \frac{1}{k}\sum_{j=i-k/2}^{i+k/2} f(x_j)]^2 = \frac{f''(x_i)}{2}\sum_{j=i-k/2}^{i+k/2}(x_j - x_i)^2 + \cdots \tag{A5}$$

From Eqs. A4 and A5, it is obvious that increasing span k (*if* $f''(x_i) \neq 0$) will increase the first term while reducing the second term of Eq. A4. Reducing k has an opposite effect.

We define the average squared residual as

$$\hat{l}_{cv}^2(k) = \frac{1}{n}\sum_{i=1}^{n}[y_i - s_{(i)}(x_i|k)]^2 \tag{A6}$$

where $s_{(i)}(x_i|k)$ is the smoothed value at x_i which is calculated from the $k-1$ observations excluding x_i, by a method referred to as 'cross-validation' or 'predictive sample reuse'.

The optimal span k_{cv} is determined by minimizing the Eq. A6 as

$$k_{cv} = \min_{0<k\le n} \hat{l}_{cv}^2(k) \tag{A7}$$

The procedure for minimization is to select several (4 to 7) discrete values of k in the range $0 < k \le n$ and calculate

$\hat{l}_{cv}^2(k)$. The k value corresponding to the smallest $\hat{l}_{cv}^2(k)$ is chosen for the optimal span.

The optimal span determined using the above method is constant over the entire data range. Since the $f''(x_i)$ and variance σ^2 usually change over the data range, it is desirable to allow the optimal span to change accordingly. This can be achieved by replacing global averaging (n) squared in Eq. A6 with a local data range (L , $0 < L \le n$). The size of L can be determined using a second level of cross-validation. In most circumstance, L is selected between (0.2-0.3) n .

To make the Supersmoother computationally efficient, several techniques are employed to minimize the computation. For local averaging at each point instead of recomputing the mean every time, an updating formula is employed which reduces the computation from nk to n .

The advantage of the Supersmoother is that it is mathematically rigorous and computationally efficient. This is credited to the use of cross-validation in selecting the optimal span.

Supersmoother for ACE. For a finite data set the ACE algorithm replaces the conditional expectation calculations by a data smoother, such as the Supersmoother discussed above. For bivarite cases, the algorithm can now be formulated as follows

For $i = 1,n$

Step 1: set $\theta^0(y_i) = y_i, \phi^0(x_i) = 0$ as initial value

Then, for $k \ge 1$

Step 2: calculate $\phi^k(x_i) = s_x[\theta^{k-1}(y_i)]$

Step 3: calculate $\theta^k(y_i) = s_y[\phi^k(x_i)]/\|s_y[\phi^k(x_i)]\|$

Step 4: iterate Steps 2 to 3 on k , until the errors (see Eq. 5) satisfy the following

$$|e^2[\theta^k(y_i),\phi^k(x_i)] - e^2[\theta^{k-1}(y_i),\phi^{k-1}(x_i)]| \le \delta \quad \text{(A8)}$$

where δ is a given tolerance. The final $\theta^k(y_i)$ and $\phi^k(x_i)$ are the solutions to the optimal transformations $\theta^*(y_i)$ and $\phi^*(x_i)$.

TABLE 1- COMPARISON OF CORRELATION R^2 FOR THE THREE TRANSFORMATION METHODS (SPRABERRY FORMATION IN PARKS FIELD)

Transformation method	Correlation R^2
Logarithmic	0.692
Normal Score	0.658
ACE	0.799

TABLE 2- COMPARISON OF CORRELATION R^2 FOR THE FIVE TRANSFORMATION METHODS (ADMIRE SAND IN THE EL DORADO FIELD)

Transformation method	Correlation R^2
Logarithmic	0.626
Normal Score	0.743
Power	0.758
Power by ACE*	0.772
ACE	0.778

* the exponents of power transformations are taken from ACE transformations.

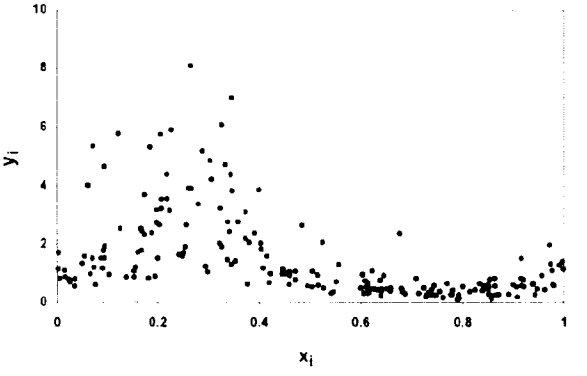

Fig. 1a-Scatterplot of y_i vs. x_i simulated from bivariate model $y_i=\exp(\sin(2\pi x_i)+\varepsilon_i/2)$, where x_i is drawn from uniform distribution U(0,1), and ε_i is independently drawn from normal distribution N(0,1).

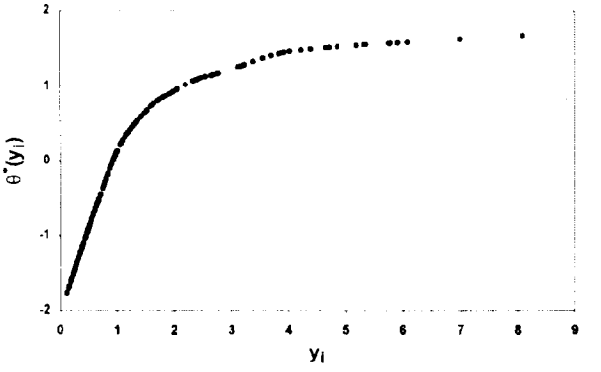

Fig. 1b-Optimal transformation of y_i by ACE.

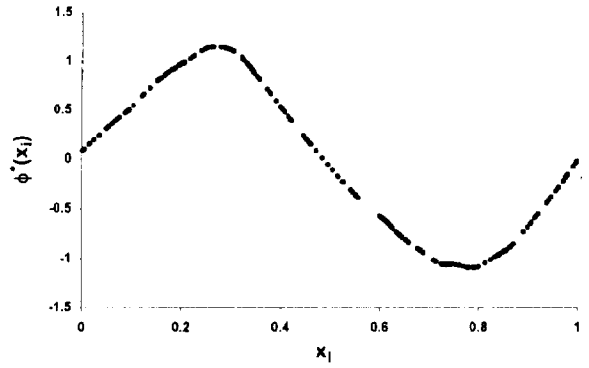

Fig. 1c-Optimal transformation of x_i by ACE.

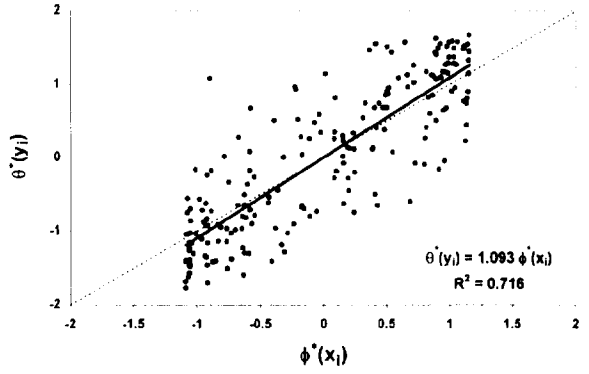

Fig. 1d-Optimal transformation of y_i vs. optimal transformation of x_i by ACE. The solid straight line represents linear regression of the data.

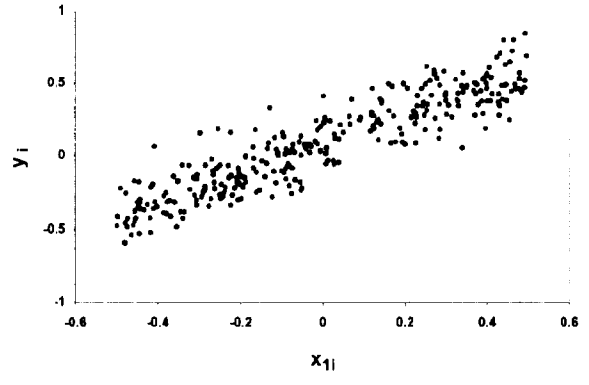

Fig. 2a-Scatterplot of y_i vs. x_{1i} simulated from multivariate model $y_i=x_{1i}+x_{2i}^2+x_{3i}^3+0.1\varepsilon_i$, where x_{1i}, x_{2i}, and x_{3i} are independently drawn from uniform distribution U(-0.5,0.5), and ε_i is independently drawn from standard normal distribution N(0,1).

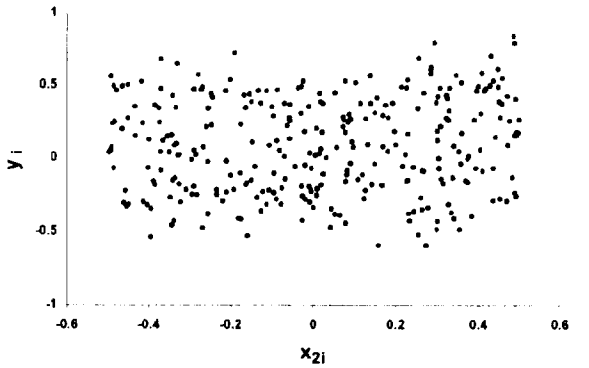

Fig. 2b-Scatterplot of y_i vs. x_{2i} simulated from multivariate model $y_i=x_{1i}+x_{2i}^2+x_{3i}^3+0.1\varepsilon_i$.

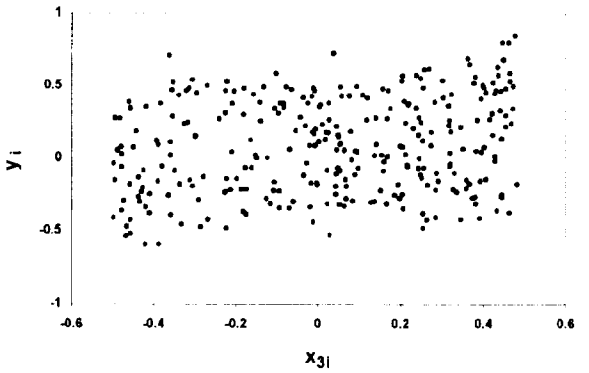

Fig. 2c-Scatterplot of y_i vs. x_{3i} simulated from multivariate model $y_i=x_{1i}+x_{2i}^2+x_{3i}^3+0.1\varepsilon_i$.

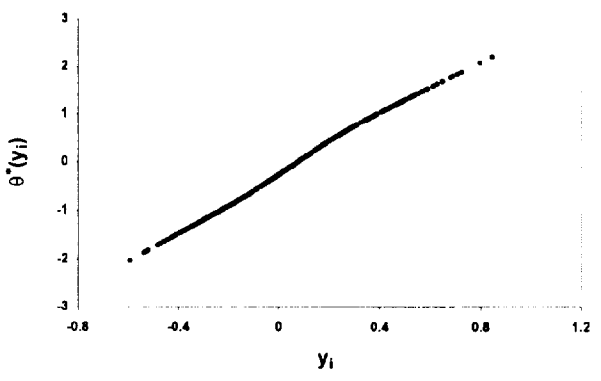

Fig. 2d-Optimal transformation of y_i by ACE.

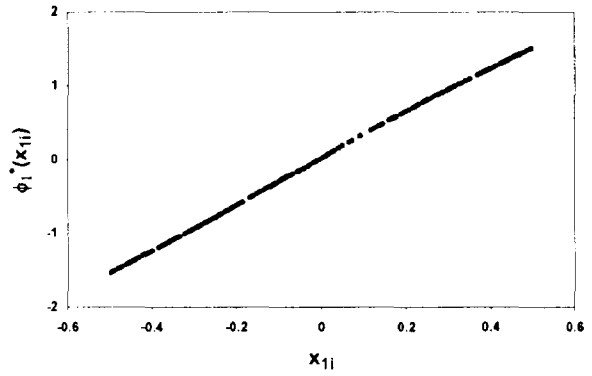

Fig. 2e-Optimal transformation of x_{1i} by ACE.

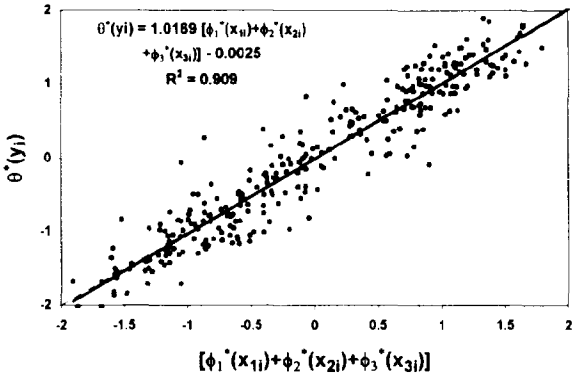

Fig. 2h-Optimal transformation of y_i vs. the sum of optimal transformations of x_{1i}, x_{2i}, and x_{3i}. The solid straight line represents a linear regression of the data.

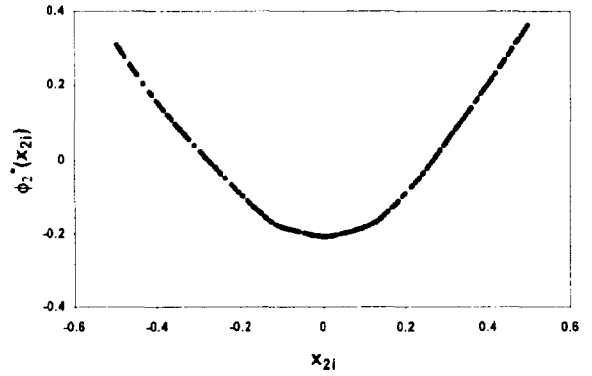

Fig. 2f-Optimal transformation of x_{2i} by ACE.

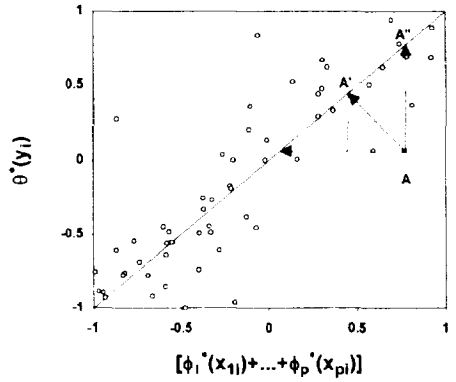

Fig. 3-Schematic drawing showing the data correction.

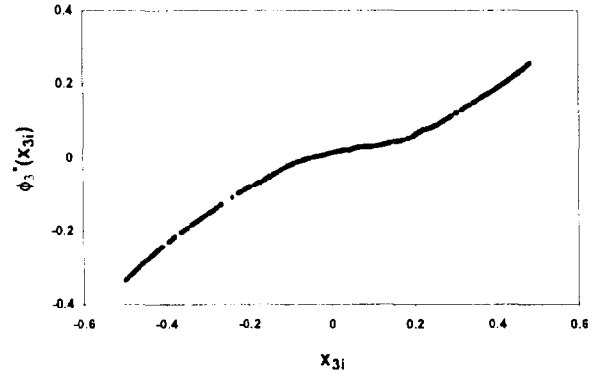

Fig. 2g-Optimal transformation of x_{3i} by ACE.

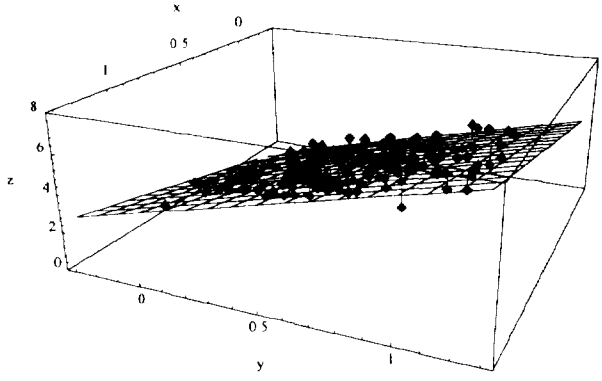

Fig. 4a-Scatterplot of z vs. x and y simulated from model z=1+2x+3y with random gaussian noise in x, y, and z.

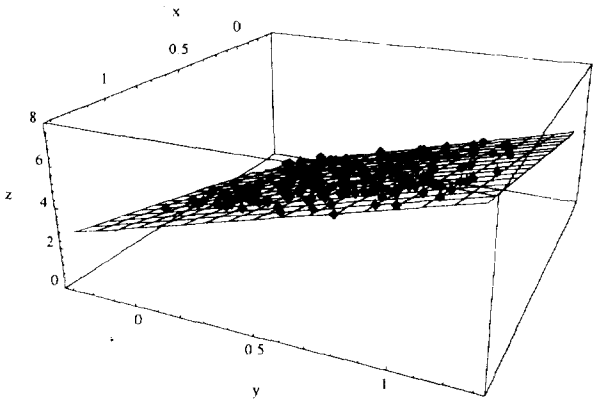

Fig. 4b-Scatterplot of z vs. x and y after data correction.

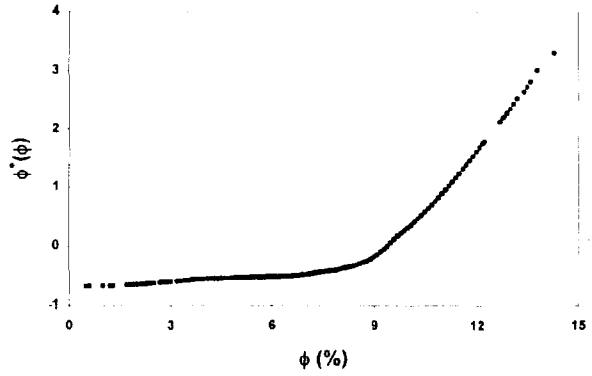

Fig. 5c-Optimal transformation of core porosity by ACE.

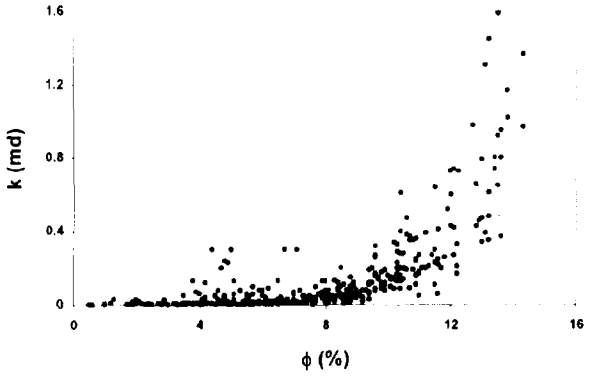

Fig. 5a-Scatterplot of core permeability vs. core porosity, Spraberry formation, Park Field.

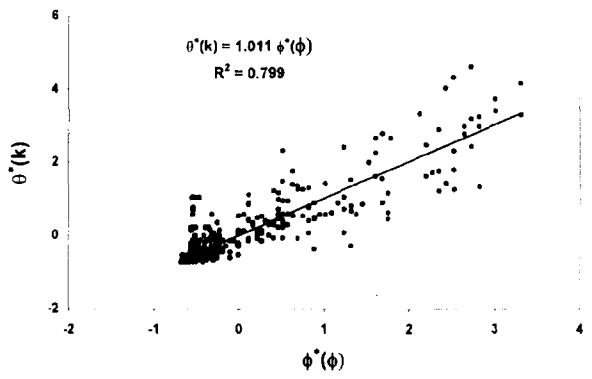

Fig. 5d-Optimal transformation of permeability vs. optimal transformation of porosity. The solid straight line represents a linear regression of the data.

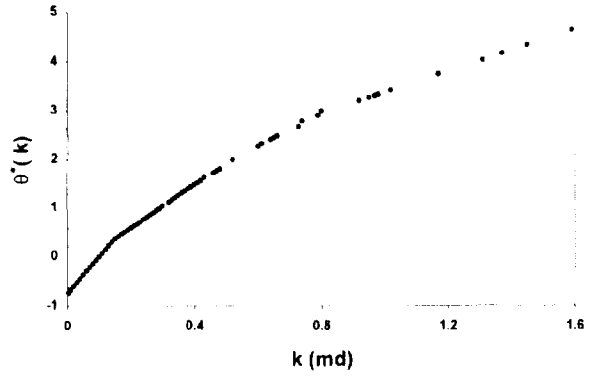

Fig. 5b-Optimal transformation of core permeability by ACE.

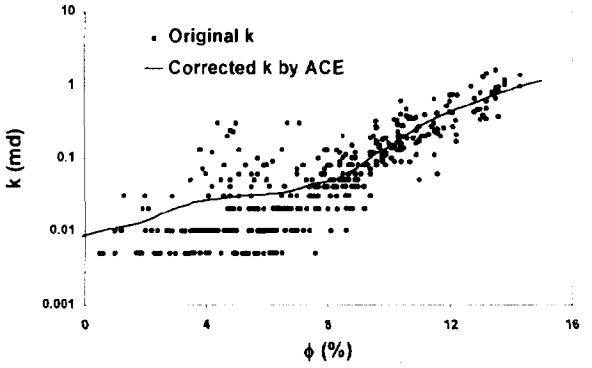

Fig. 5e-Original permeability and corrected permeability by ACE.

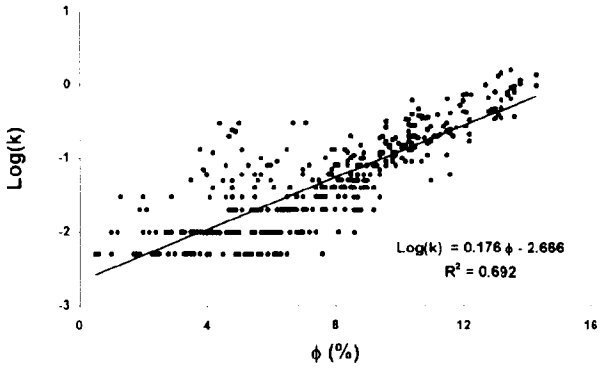

Fig. 5f-Logarithmically transformed permeability vs. porosity. The solid straight line represents a linear regression of the data.

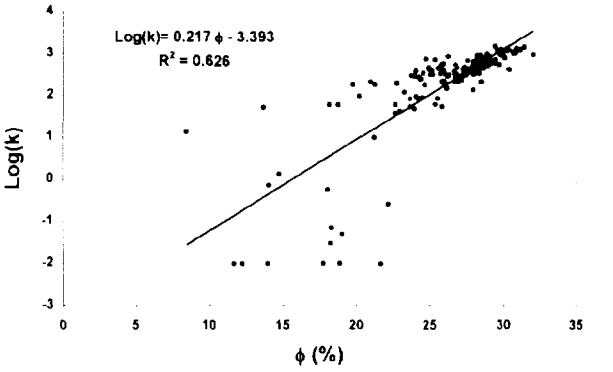

Fig. 6b-Logarithmically transformed permeability vs. porosity. The solid straight line represents a linear regression of the data.

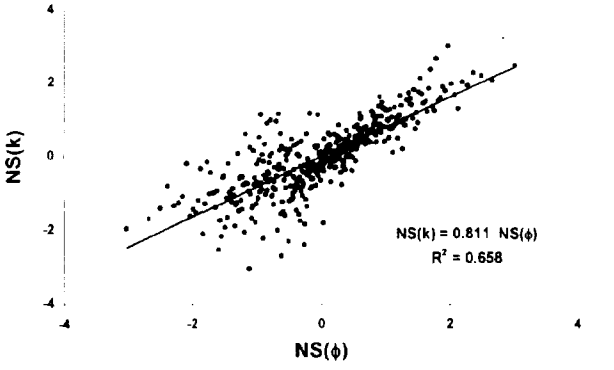

Fig. 5g-Normal score transformed permeability vs. normal score transformed porosity. The solid straight line represents a linear regression of the data.

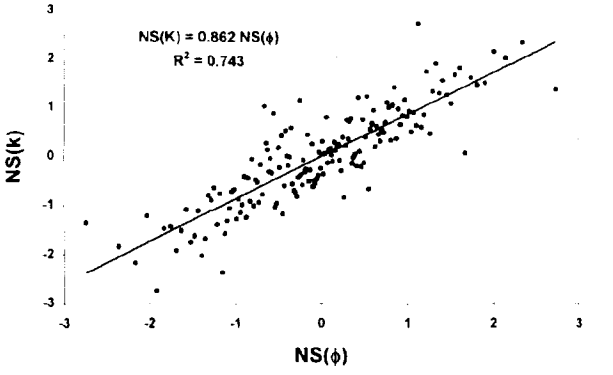

Fig. 6c-Normal score transformed permeability vs. normal score transformed porosity. The solid straight line represents a linear regression of the data.

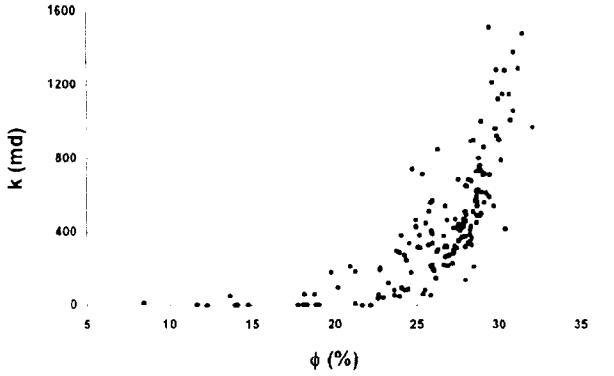

Fig. 6a-Scatterplot of core permeability vs. core porosity, the Admire sand, the El Dorado Field.

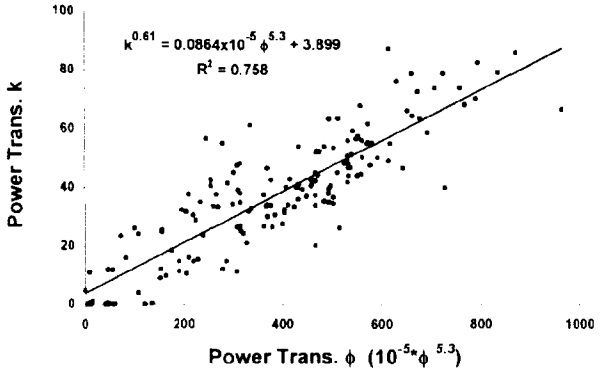

Fig. 6d-Power transformed permeability vs. power transformed porosity. The solid straight line represents a linear regression of the data.

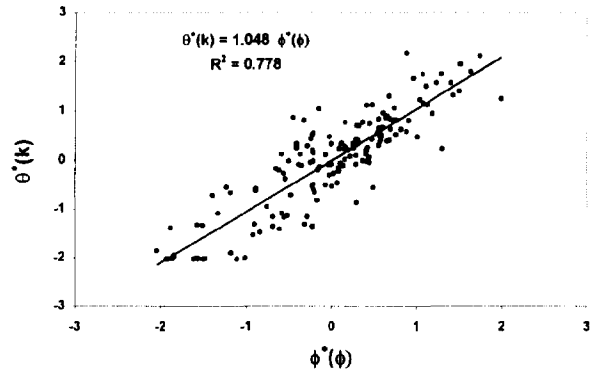

Fig. 6e-Transformed permeability vs. Transformed porosity by ACE. The solid straight line represents a linear regression of the data.

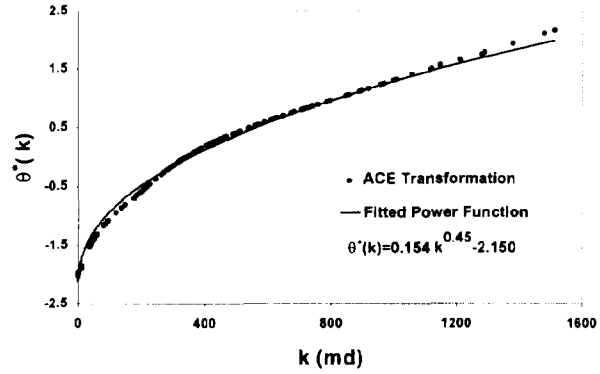

Fig. 6f-Optimal transformation of permeability by ACE. The solid line represents a fitted power function.

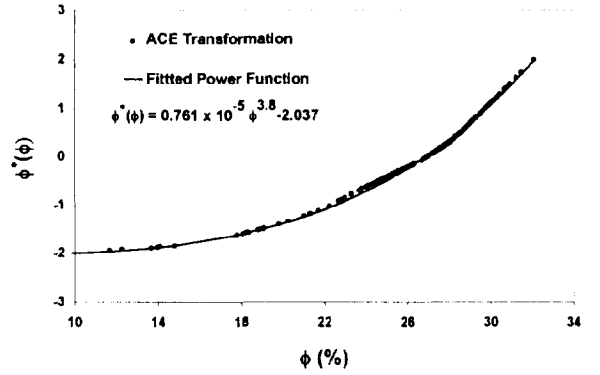

Fig. 6g-Optimal transformation of porosity by ACE. The solid line represents a fitted power function.

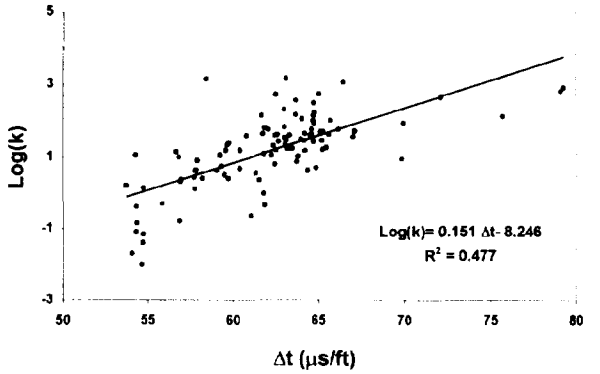

Fig. 7a-Logarithmically transformed permeability vs. sonic travel time, Schneider Buda Field. The solid straight line represents a linear regression of the data.

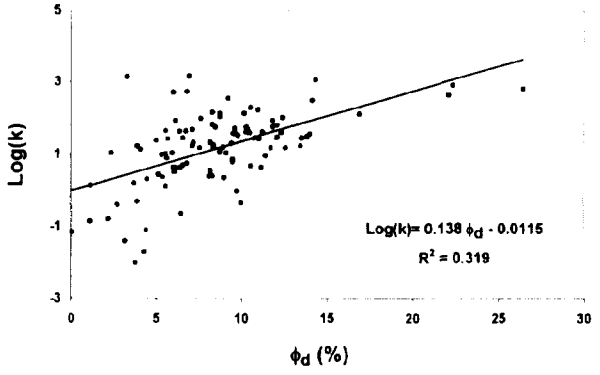

Fig. 7b-Logarithmically transformed permeability vs. density derived porosity. The solid straight line represents a linear regression of the data.

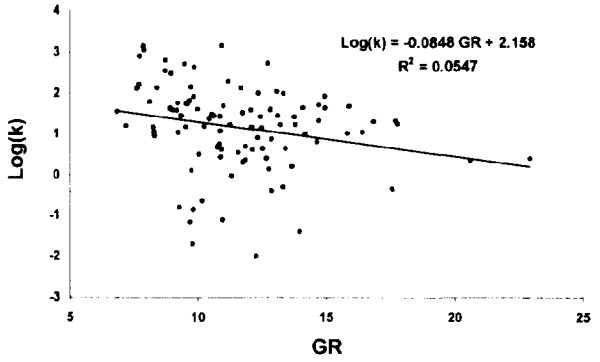

Fig. 7c-Logarithmically transformed permeability vs. gamma ray. The solid straight line represents a linear regression of the data.

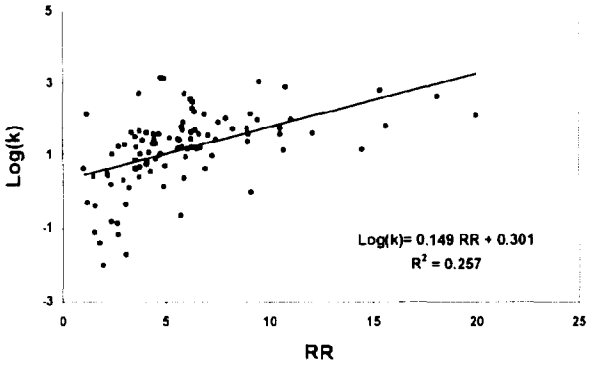

Fig. 7d-Logarithmically transformed permeability vs. resistivity ratio. The solid straight line represents a linear regression of the data.

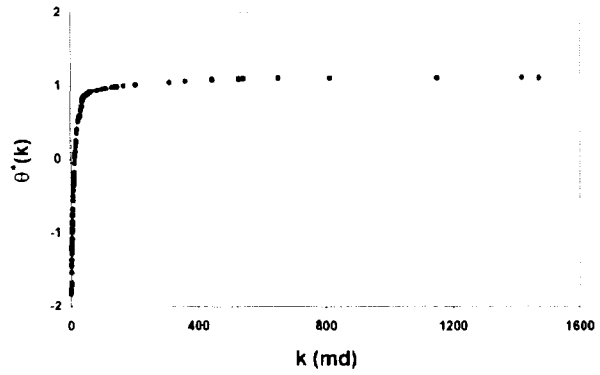

Fig. 7e-Optimal transformation of permeability by ACE.

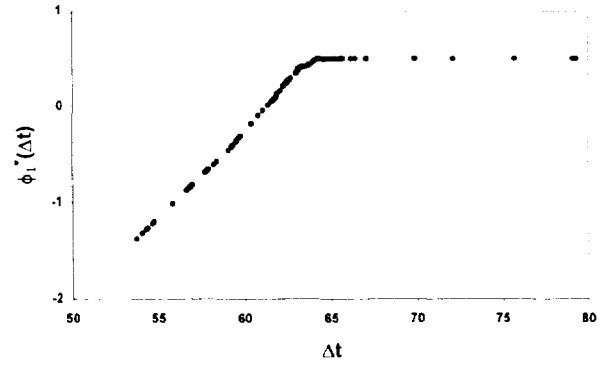

Fig. 7f-Optimal transformation of sonic travel time by ACE.

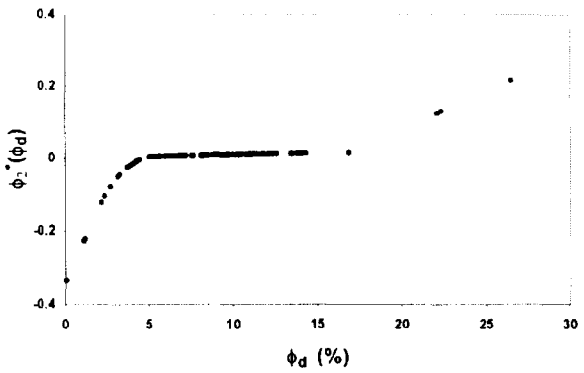

Fig. 7g-Optimal transformation of density derived porosity by ACE.

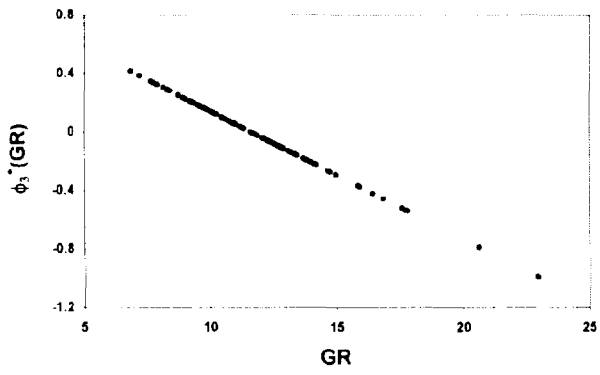

Fig. 7h-Optimal transformation of gamma ray by ACE.

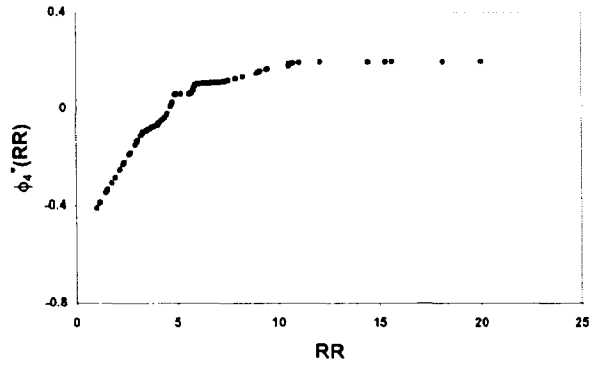

Fig. 7i-Optimal transformation of resistivity ratio by ACE.

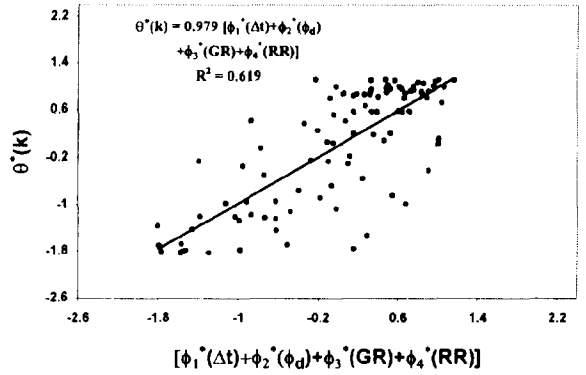

Fig. 7j-Optimal transformation of permeability vs. the sum of optimal transformation of well log variables. The solid straight line represents a linear regression of the data.

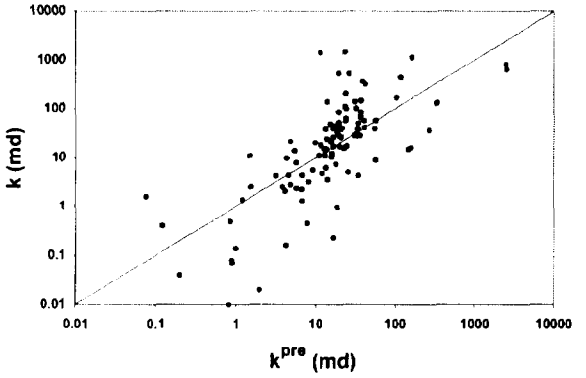

Fig. 7k-Core permeability vs. the permeability predicted from well logs using ACE.