OVERVIEW OF STATISTICAL METHODS IN LOG ANALYSIS

by

John C. Davis
John H. Doveton

Presented at SPWLA Computer Applications Workshop
June 28, 1990
Lafayette, LA

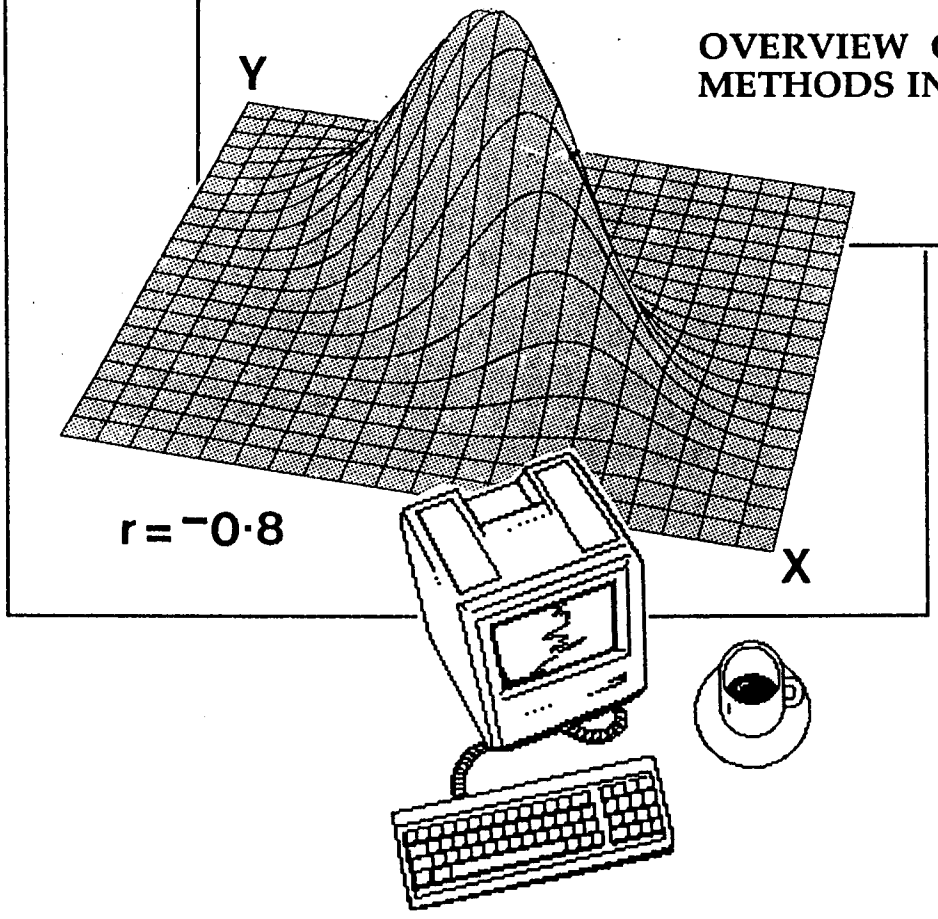
Kansas Geological Survey
Open-file Report
90-26

Disclaimer

SPWLA COMPUTER APPLICATIONS WORKSHOP
June 28, 1990 Lafayette, La

JOHN C. DAVIS AND JOHN H. DOVETON
Kansas Geological Survey
Lawrence, Ks

OVERVIEW OF STATISTICAL METHODS IN LOG ANALYSIS



Methods of Fitting a Straight Line

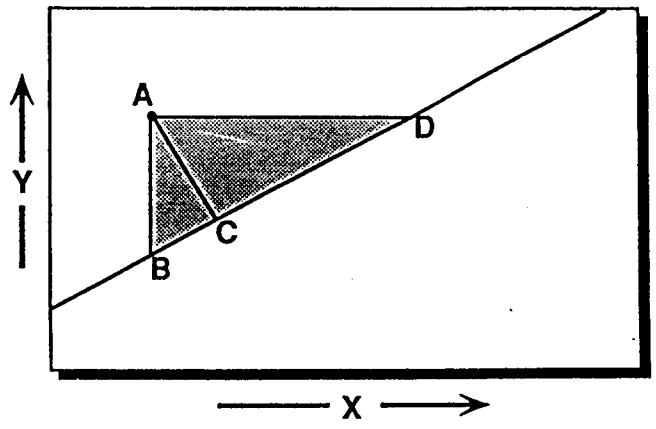
Calculation of a line of best fit by a mathematical method that minimizes the deviations of the points from the line in some manner.

A. Ordinary Least Squares—Minimizes the sum of the squared deviations of Y from the fitted line (minimizes lines A—B).

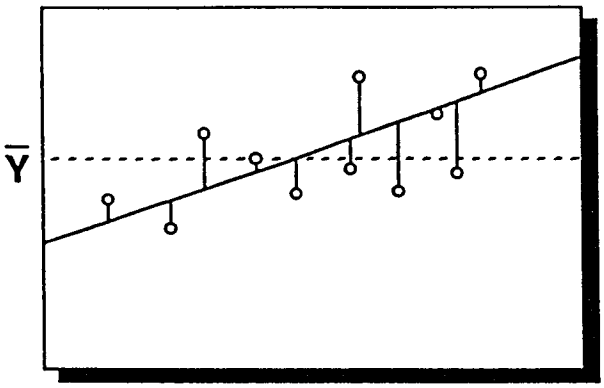
B. Inverse Least Squares—Minimizes the sum of the squared deviations of X from the fitted line (minimizes lines A—D).

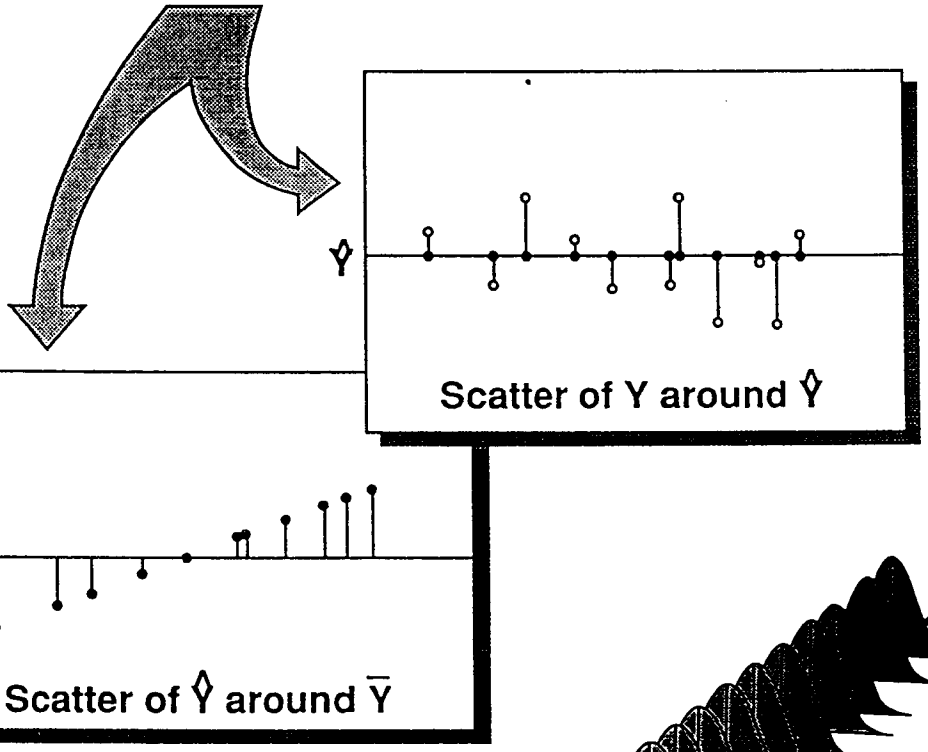
C. Least Normal Squares or Major Axis—Minimizes the sum of the squared perpendicular deviations of the points from the fitted line (minimizes lines A—C).

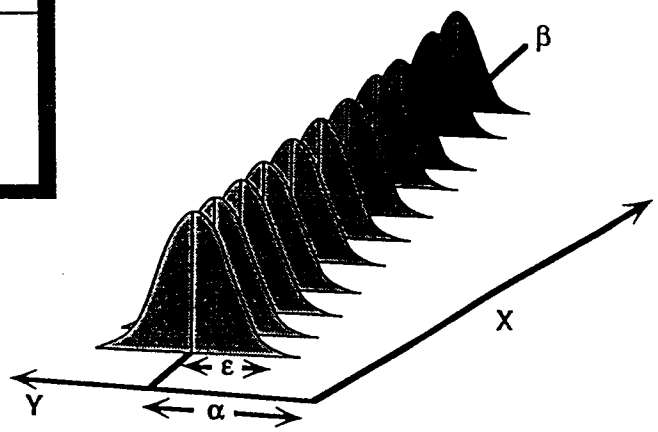
D. Reduced Major Axis (RMA)—Minimizes the sum of areas of the right triangles formed between the data points and the fitted line (minimizes areas ABD).



Regression divides the variance of Y into two parts: That due to the regression, and that due to deviations from the regression.



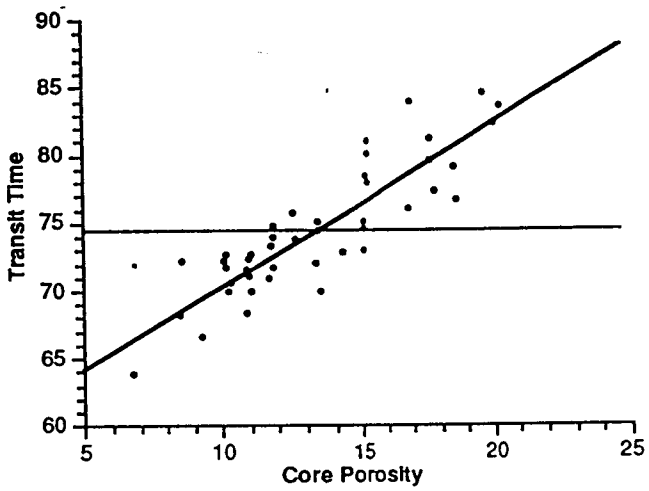




$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	*F* Test
Linear Regression	SS_R	1	MS_R	MS_R/MS_D
Deviation	SS_D	$n - 2$	MS_D	
Total Variation	SS_T	$n - 1$		

SIMPLE REGRESSION EXAMPLE: Regression of sonic transit time on core porosity in the Sadlerochit of the Prudhoe Bay field. If there is a strong relationship between the two, then perhaps the sonic logs can be calibrated against core porosity and used to estimate porosity in non-cored wells.



$N = 44$ $S_y = 4.712$ $S_x = 3.392$
$S_{xy} = 13.985$ $r_{xy} = 0.875$

The regression line is

$$\Delta t = 58.056 + 1.216\phi$$

The regression line is shown on the scatter plot, as is a line representing the mean $\overline{Y}$. Are the two significantly different?

The significance of the fitted regression can checked by analysis of variance. The necessary ANOVA table is:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-test
Linear Regression	731.549	1	731.549	137.67**
Deviation	223.172	42	5.314	
Total	954.721	43		$F_{1,42} = 4.07$

The computed test value greatly exceeds the critical value for a significance level of 95%, and even exceeds 99%. The relationship between core porosity and sonic log transit time is highly significant.

The regression

$$\Delta_t = 58.056 + 1.216\phi$$

will allow us to predict sonic log transit times from known core porosities. Unfortunately, what is wanted is the reverse! This reverse relationship can be found in two ways.

1. Inverse regression

By reversing X and Y, we can find the regression

$$\phi = -33.435 + 0.6302\Delta_t$$

Unfortunately, the uncertainty is in the core porosity for a given sonic transit time, yet the core porosities are presumed to be "true" values!

2. Calibration

The coefficients of the regression of transit times on core porosities can be determined and then the equation inverted to solve for estimated values of porosity. The inverse equation is

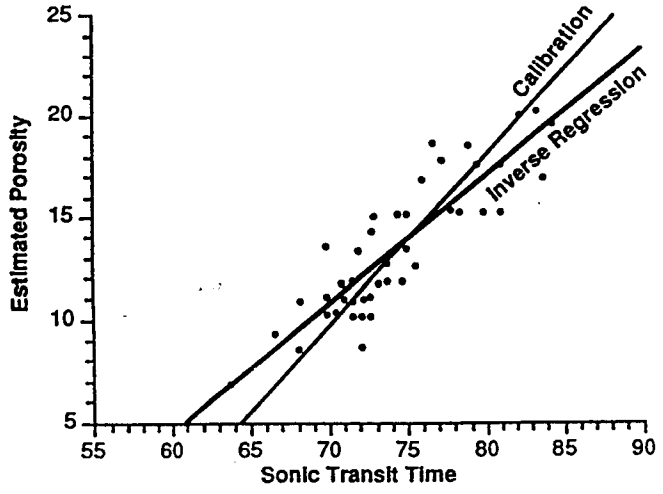
$$\hat{X} = (Y - a)/b$$

Substituting in the regression coefficients gives

$$\hat{\phi} = (\Delta_t - 58.056)/1.216$$

Since the coefficients of Y = a + bX and X = a + bY are not the same, the two approaches give different answers.

For values near the means, both inverse regression and calibration procedures give similar answers. But for predictions beyond the range of the data used to calibrate the predicting equation (beyond the range of known core porosities), the calibration procedure gives much better results. Calibration is advocated by most (but not all) statisticians.



The differences that result from fitting a regression by alternative procedures can be demonstrated using the 44 measurements of core porosity and sonic transit time data from Sadlerochit Zone 3. Here, X = core porosity, ϕ and Y = sonic well log transit time, Δ_t.

Ordinary least squares (Y on X):

$$Y = a + bX \rightarrow \Delta_t = 58.056 + 1.216\phi$$

Inverse least squares (X on Y):

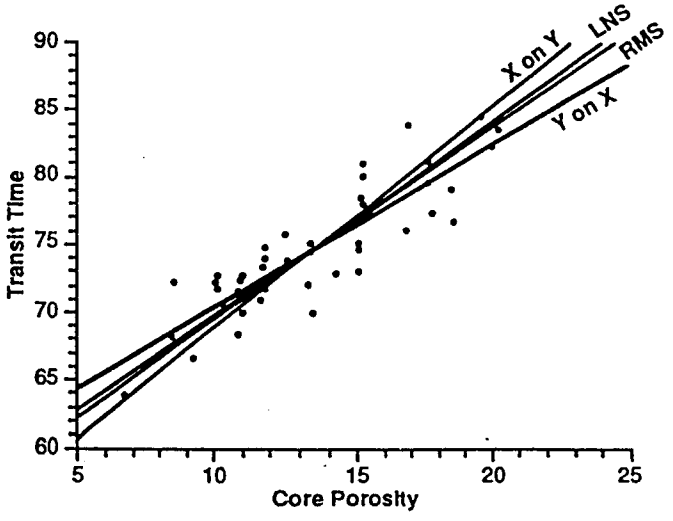
$$X = a + bY \rightarrow \phi = -33.435 + 0.631\Delta_t$$

Least normal squares (major axis):

$$Y = a + bX \rightarrow \Delta_t = 54.862 + 1.453\phi$$

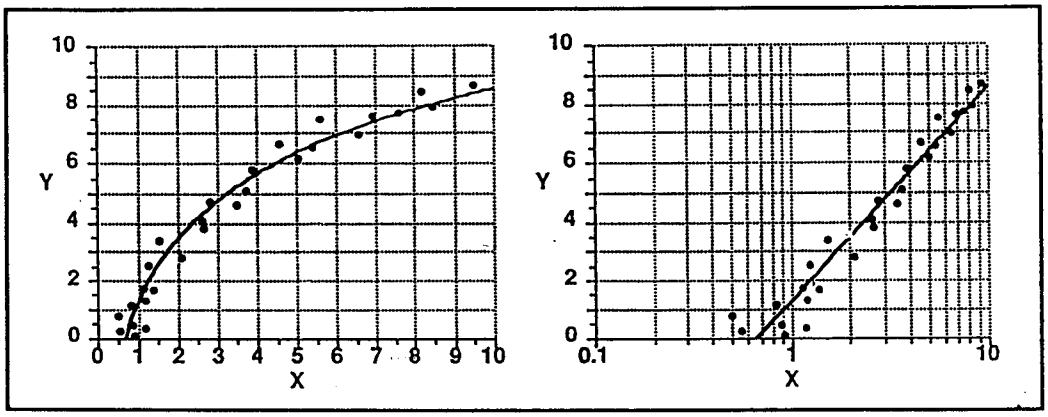
Reduced major axis:

$$Y = a + bX \rightarrow \Delta_t = 55.724 + 1.389\phi$$



LOGARITHMIC TRANSFORMATIONS

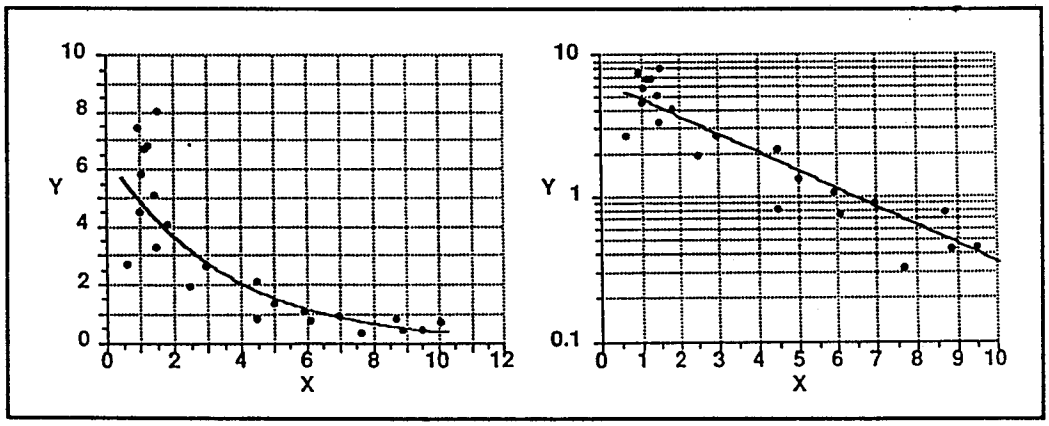
Regressing Y on the natural log of X : $Y = a + b \ln X$
This transformation will not affect the distribution of the deviations of Y from the fitted regression.



Regressing the natural log of Y on X: $\ln Y = a + b X$
fits an exponential model :

$$Y = e^{a} + b X$$

The deviations of Y from the model will not be symmetric, but the deviations of ln Y will be.



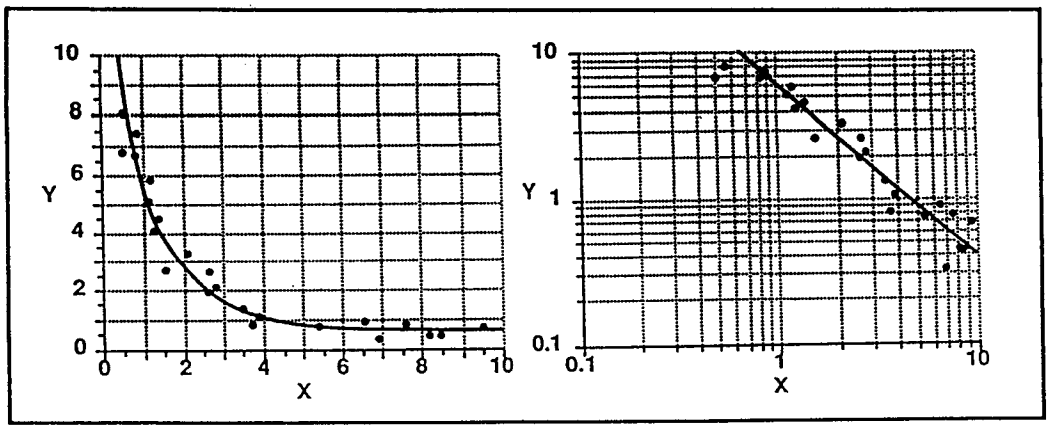
Regressing the natural log of Y on the natural log of X :

$$\ln Y = a + b \ln X$$

fits a multiplicative model :

$$Y = a X^{b}$$

Because of the logarithmic transformation of Y, the deviations of Y from the regression will not be symmetric, although the deviations of ln Y will be.



Weighted regression

The normal equations for linear regression can be extended to allow individual observations to be weighted so their importance in determining the line of best fit is increased or decreased.

$$b = \frac{\sum w_i \sum w_i X_i Y_i - (\sum w_i X_i)(\sum w_i Y_i)}{\sum w_i \sum w_i X_i^2 - (\sum w_i X_i)^2}$$

$$a = \frac{\sum w_i Y_i - b \sum w_i X_i}{\sum w_i}$$

The weighting factors w_i can be normalized so their sum is equal to the number of observations

$$\hat{w}_i = \frac{N w_i}{\sum w_i}$$

Normalizing is not necessary to compute the regression, but it will simplify comparisons between alternative weightings of the same data and also simplifies the normal equations because then $\sum \hat{w}_i = N$.

THE GENERAL REGRESSION MODEL

A dependent (or predicted) variable Y, is regressed on m independent (predictor) variables $X_1, X_2, \ldots\ldots X_m$.
The n observation sets can be symbolized as :

DATA INPUT →

$$\begin{matrix} Y_1 & X_{11} & X_{12} & \cdots\cdots & X_{1m} \\ Y_2 & X_{21} & X_{22} & \cdots\cdots & X_{2m} \\ y_3 & \cdots\cdots & & & \\ \cdots\cdots & & & & \\ \cdots\cdots & & & & \\ Y_n & X_{m1} & X_{n2} & \cdots\cdots & X_{nm} \end{matrix}$$

n sets of observations

dependent variable | m independent variables

The regression equation is : $\hat{Y} = a_0 + a_1X_1 + a_2X_2 + \ldots\ldots + a_mX_m$
The vector of predicted values of Y for all n observation sets can be written in matrix form as :

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \hat{Y}_m \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots\cdots & X_{1m} \\ 1 & X_{21} & X_{22} & \cdots\cdots & X_{2m} \\ \cdots\cdots & & & & \\ \cdots\cdots & & & & \\ \cdots\cdots & & & & \\ 1 & X_{m1} & X_{n2} & \cdots\cdots & X_{nm} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ a_m \end{bmatrix}$$

which can be symbolized as $\hat{Y} = XA$

Now, the solution is found by minimizing the sum of squares deviations between $\hat{Y}_i$ and Y_i, given by :

$$G = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - (a_0 + a_1X_{1i} + \ldots\ldots + a_mX_{mi}))^2$$

The partial differentials : $\frac{\partial G}{\partial a_0} = 0 \ldots \frac{\partial G}{\partial a_1} = 0 \ldots\ldots \frac{\partial G}{\partial a_m} = 0$

These m equations rearranged in matrix form are :

$$\begin{bmatrix} n & \sum X_1 & \sum X_2 & \cdots & \cdots & \sum X_m \\ \sum X_1 & \sum X_1^2 & \sum X_1X_2 & \cdots & \cdots & \sum X_1X_m \\ \sum X_2 & \sum X_1X_2 & \sum X_2^2 & \cdots & \cdots & \sum X_2X_m \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum X_m & \cdots & \cdots & \cdots & \cdots & \sum X_m^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \cdots \\ \cdots \\ \cdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum X_1Y \\ \sum X_2Y \\ \cdots \\ \cdots \\ \sum X_mY \end{bmatrix}$$

$$SA = P$$

$$\therefore A = S^{-1}P$$

$$But \cdots S = X^TX \cdots and \cdots P = X^TY$$

$$\therefore A = (X^TX)^{-1}X^TY$$

which gives the coefficient unknowns for the general regression equation : $\hat{Y} = a_0 + a_1X_1 + a_2X_2 + \cdots\cdots a_mX_m$

When there is only one independent variable, X1, this is the solution for SIMPLE LINEAR REGRESSION :

$$\hat{Y} = a_0 + a_1X$$

When there are several independent variables, this is the solution for MULTIPLE REGRESSION :

$$\hat{Y} = a_0 + a_1X_1 + a_2X_2 + \cdots\cdots a_mX_m$$

When the independent variables are powers of a single independent variable, this is the solution for POLYNOMIAL REGRESSION : $\hat{Y} = a_0 + a_1X + a_2X^2 + \cdots\cdots a_mX^m$

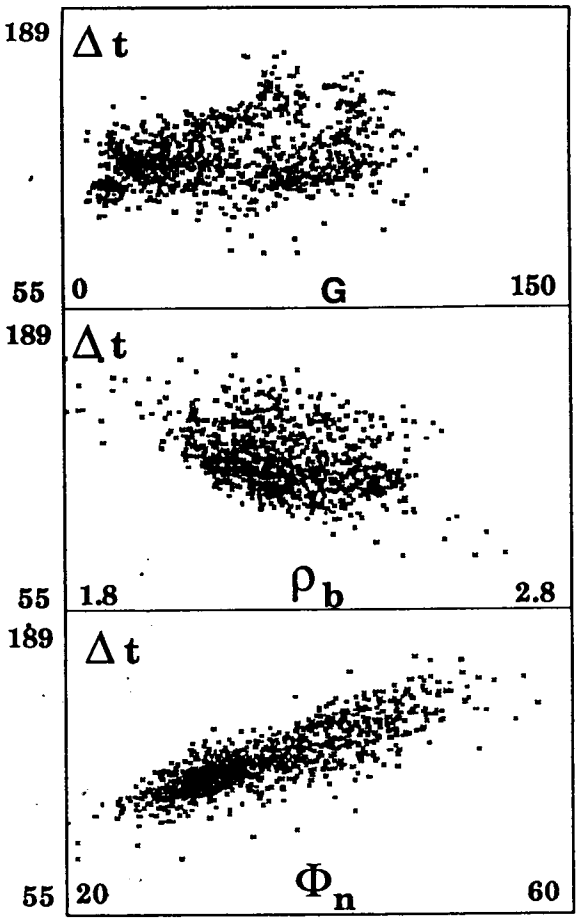
When Y is measured at geographic locations and two independent variables are polynomial combinations of geographic coordinates, this is the solution for TREND SURFACE ANALYSIS : $\hat{Y} = a_0 + a_1U + a_2V + \cdots\cdots$

When the relationship between dependent and independent variables is of the form :

$$\hat{Y} = aX^b \cdots then \cdots \log \hat{Y} = \log a + b \cdot \log X$$

and this is a solution for NON-LINEAR REGRESSION.

MULTIPLE REGRESSION EXAMPLE : Development of equation to predict acoustic transit time, based on regression on gamma-ray, density and neutron logs. The result is used to compute a pseudo-sonic log in an equivalent section where no sonic tool was run, for the purpose of generating a synthetic seismogram as reference for field records.

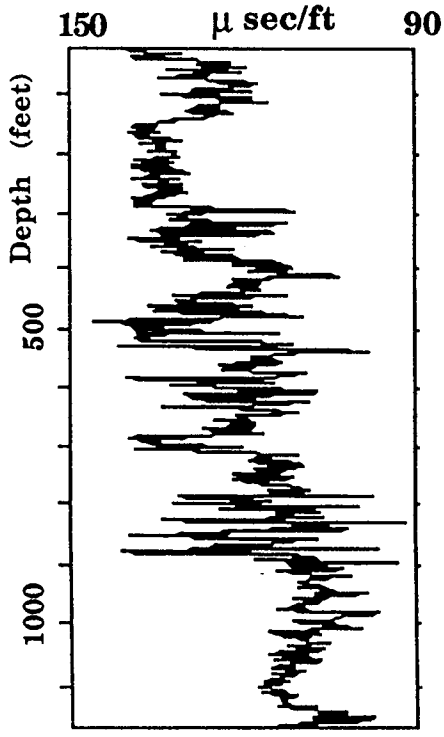


standardized partial regression cofficients :

0.05 -0.06 0.81

$$\hat{\Delta t} = 77.2 + 0.03G - 7.4\rho_b + 1.69\Phi_n$$

pseudosonic log



Fort Hays Ls

Blue Hill Sh

Greenhorn Ls

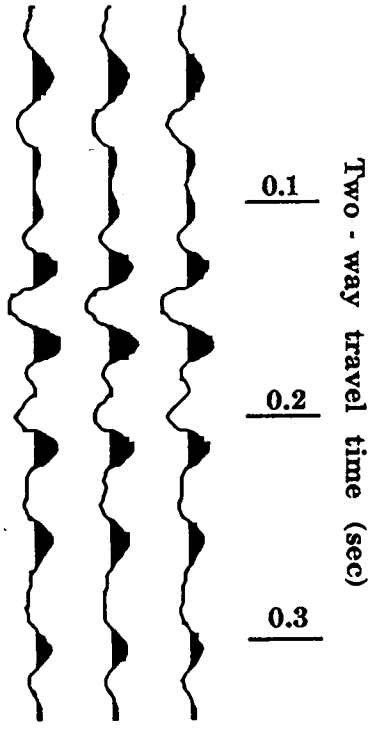
Graneros Sh

Dakota

Longford/Cheyenne

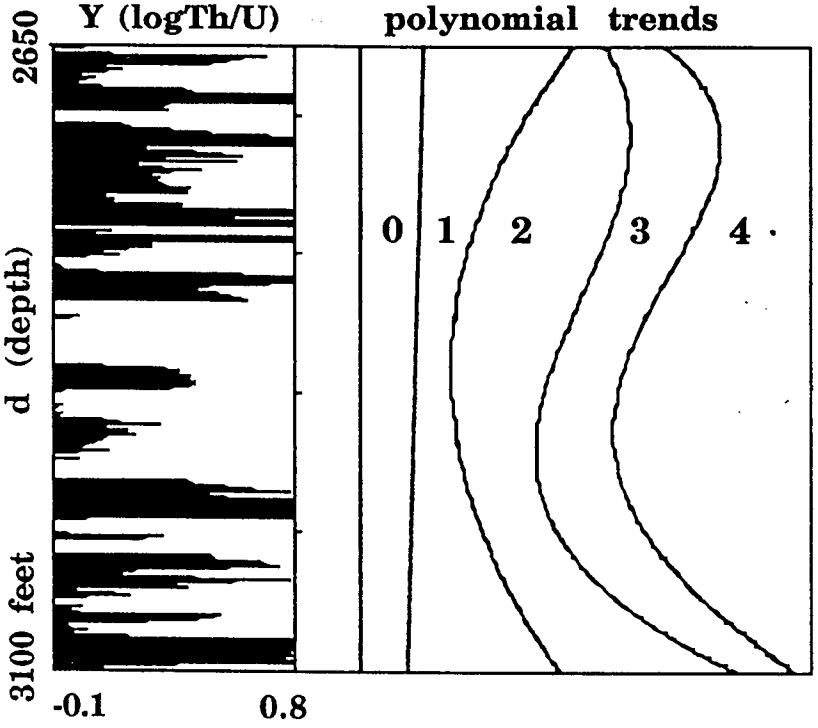
Cedar Hills Ss

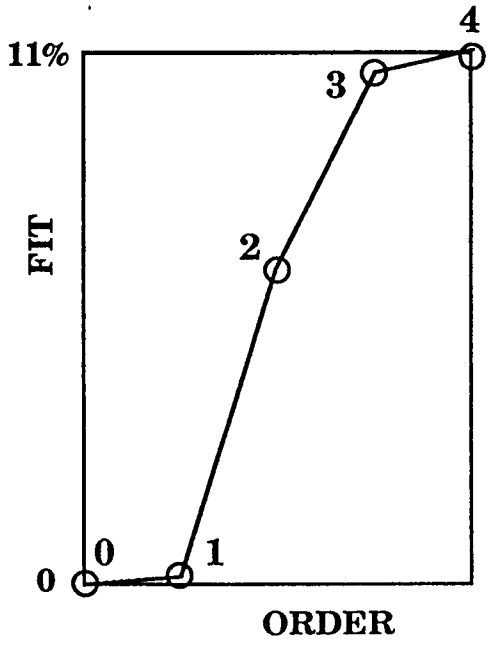
synthetic seismogram (Ricker 30 Hz)

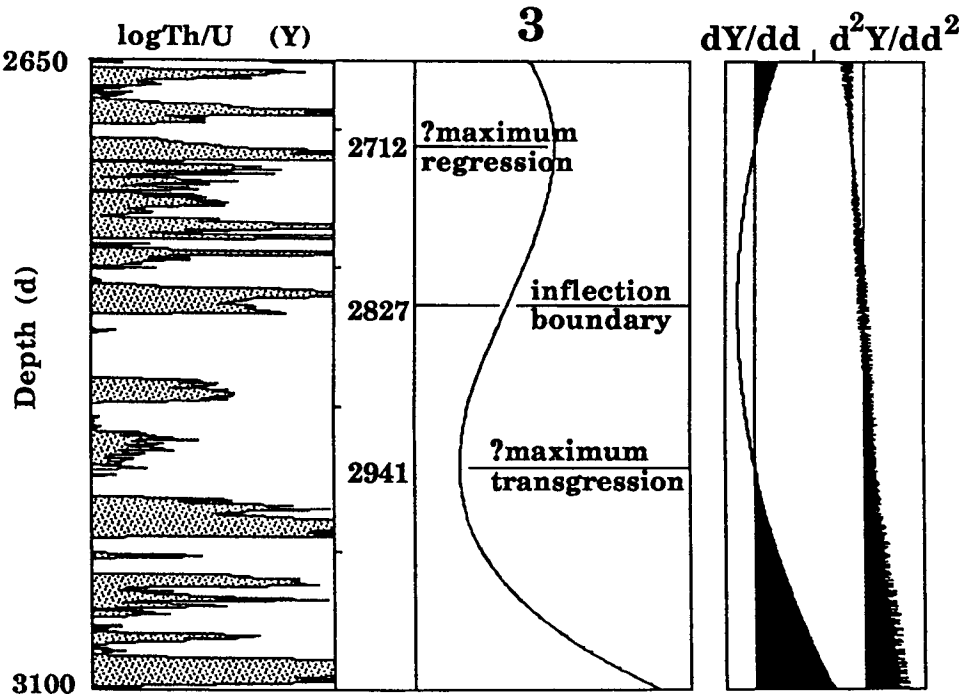


POLYNOMIAL REGRESSION EXAMPLE :
Regression of log(Th/U) on depth in the Permian Chase Group as indicator of long-term trends in redox potential and transgression/regression history

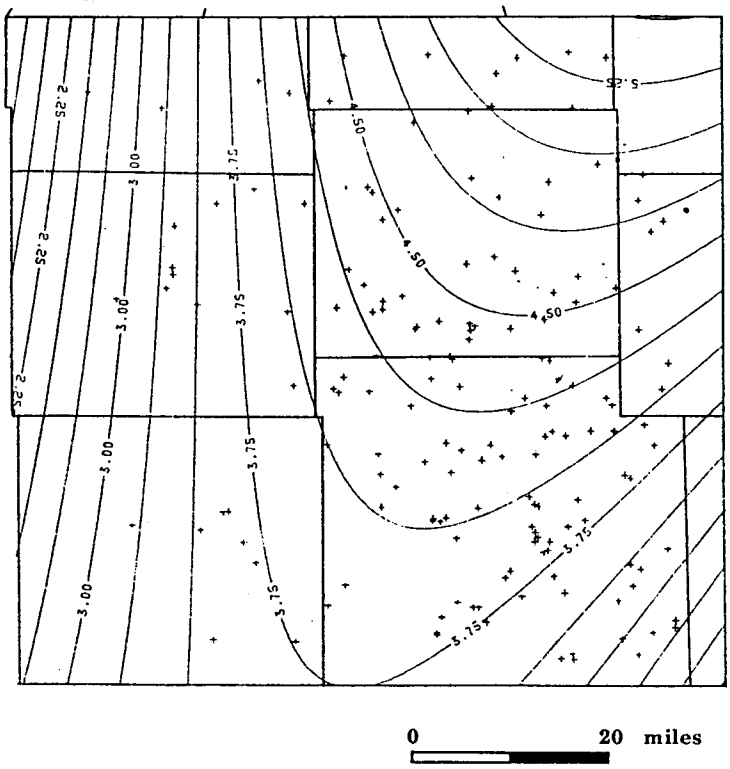
$$\hat{Y}\ (\log Th/U) = a_0 + a_1 d + a_2 d^2 + \ldots\ldots + a_m d^m$$

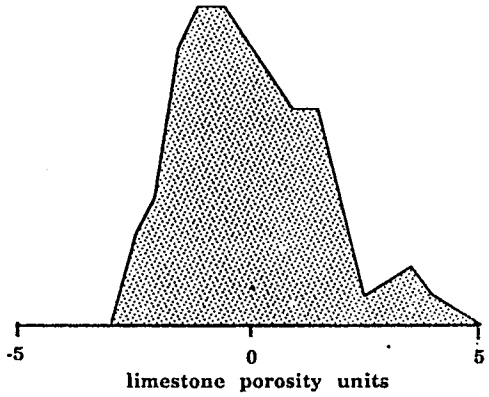






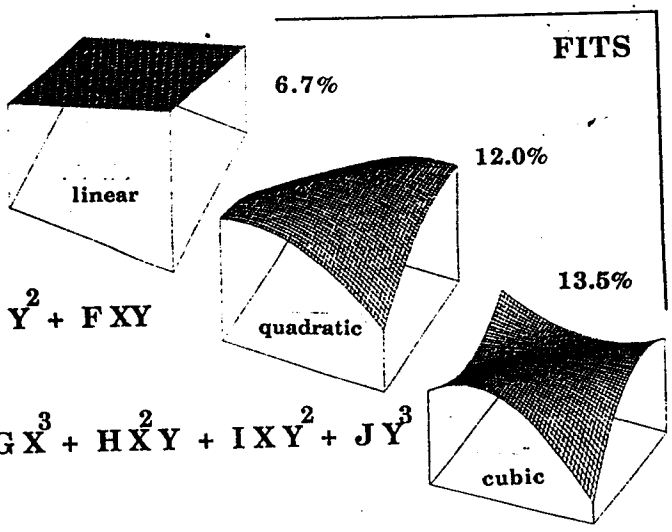
TREND SURFACE ANALYSIS EXAMPLE :
Fit of regression surfaces to the neutron porosity of the "Lower Limestone" zone of the Viola in south-central Kansas, as polynomial functions of the geographic well coordinates, X and Y. The trends pick up major regional changes in neutron porosity, while the residuals for this zone are mostly linked with tool error. The procedure is a useful method for normalization of log data.



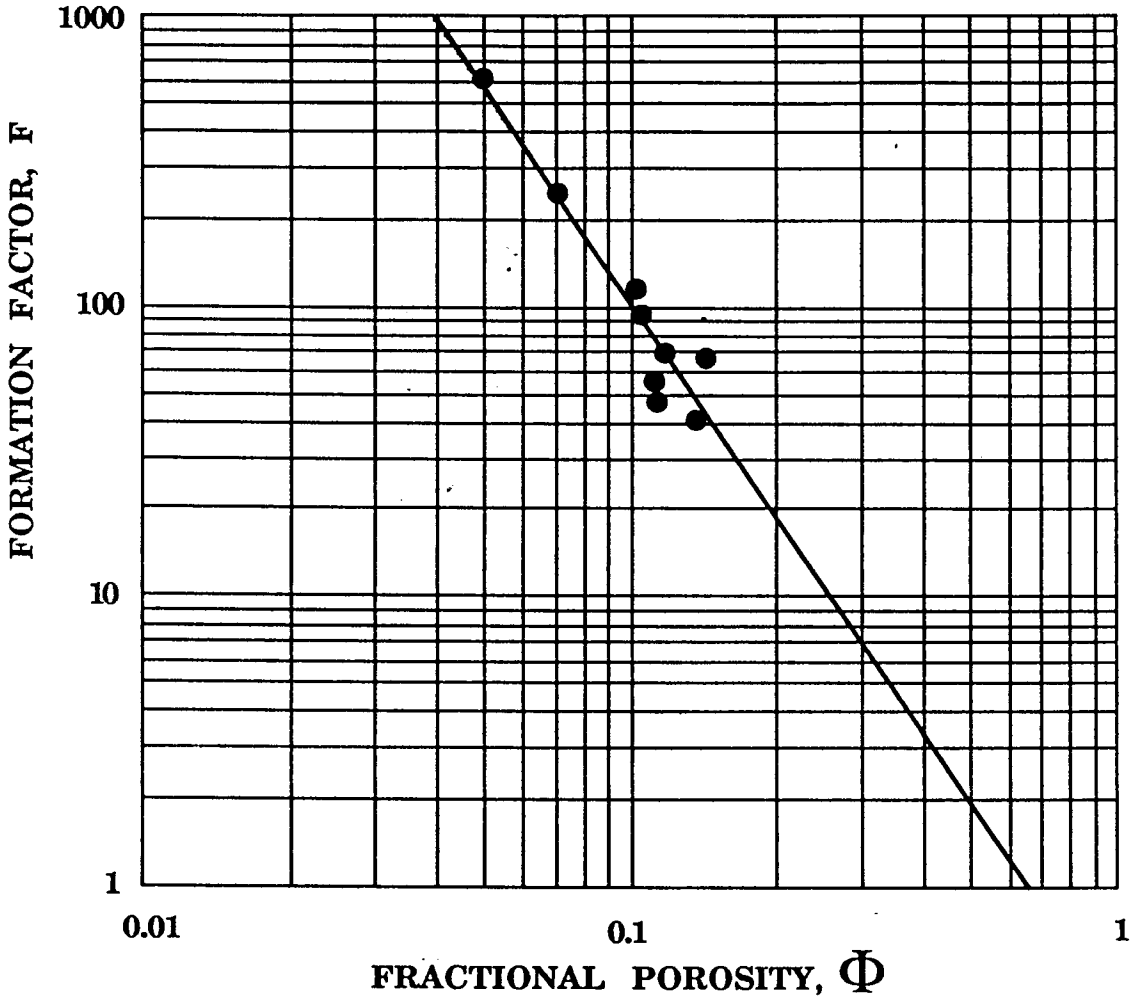


Quadratic trend-surface neutron porosity residuals

$$\hat{\Phi}_n = A + BX + CY + DX^2 + EY^2 + FXY + GX^3 + HX^2Y + IXY^2 + JY^3$$



NON - LINEAR REGRESSION EXAMPLE :
Calculation of Archie equation constants for the Arbuckle Limestone, based on core measurements of formation factor and porosity



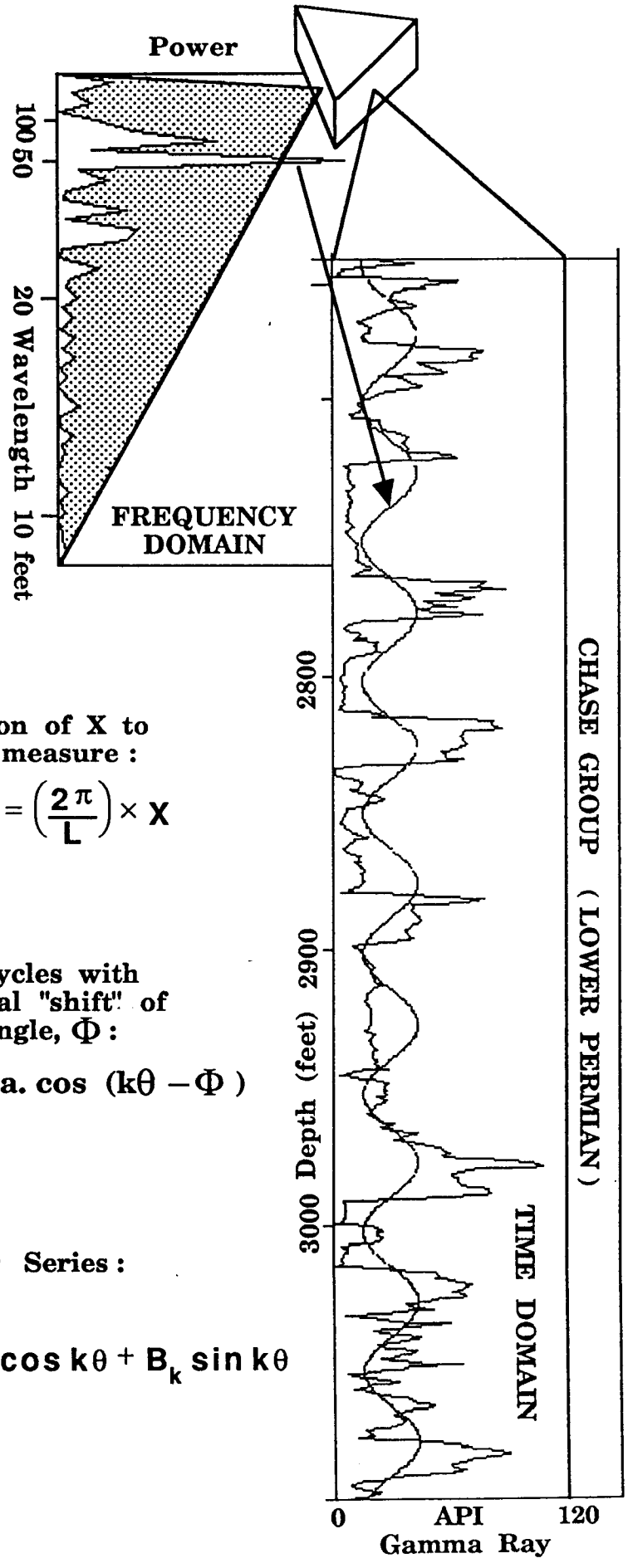
Equation of regression line (F - on - Φ) :

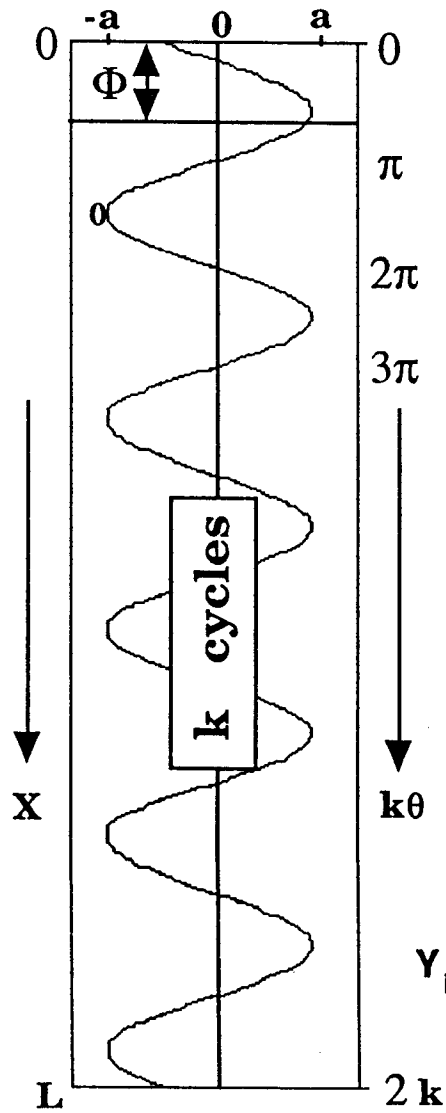
$$\log \hat{F} = -0.445 - 2.444 \log \Phi$$

$$\therefore \hat{F} = \frac{0.36}{\Phi^{2.44}}$$

(The reduced major axis (RMA) solution is : $F = \frac{0.27}{\Phi^{2.57}}$)

DISCRETE FOURIER ANALYSIS EXAMPLE :
Evaluation of potential cyclicity in the carbonate - shale alternations in the Permian Chase Group, through harmonic analysis of the gamma ray log. The pronounced spectral peak accounts for 22% of the total power (variance) and corresponds to a wavelength of about 50 feet.





Conversion of X to angular measure :

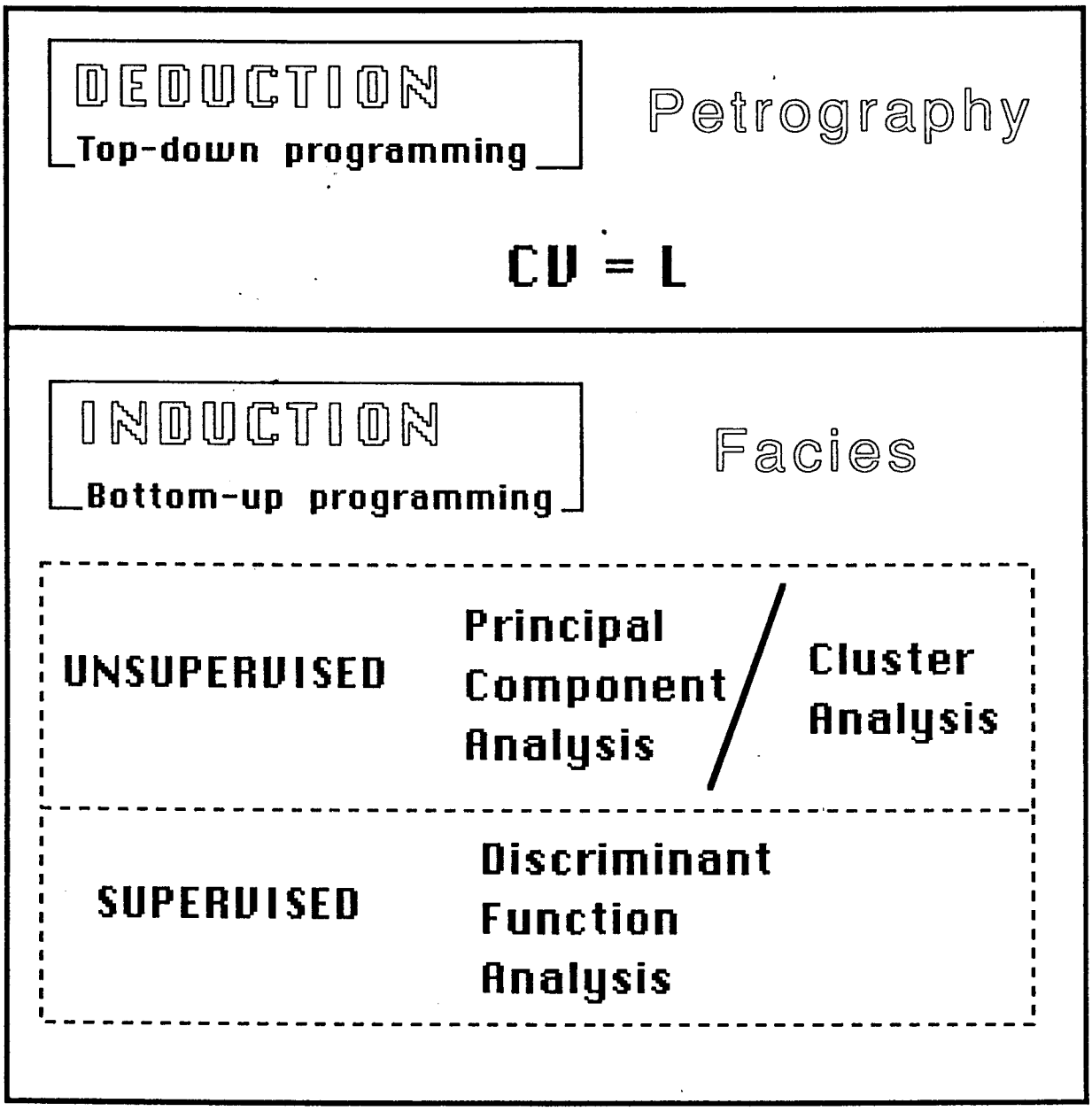
$$\theta = \left(\frac{2\pi}{L}\right) \times X$$

For k cycles with an initial "shift" of phase angle, Φ :

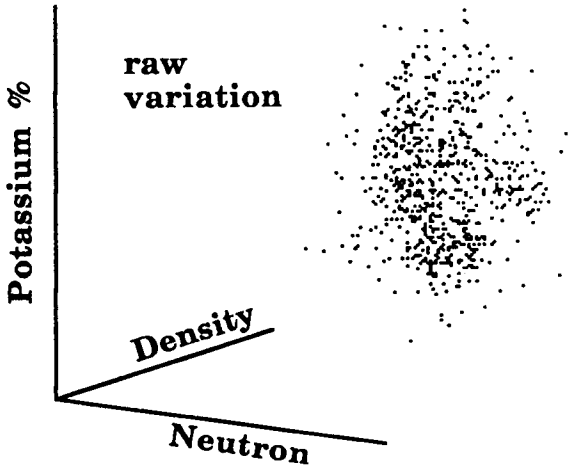
$$Y = a.\cos\,(k\theta - \Phi)$$

Fourier Series :

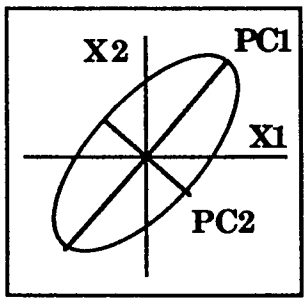
$$Y_l = \sum_{k=0}^{n/2} A_k \cos k\theta + B_k \sin k\theta$$


DEDUCTION
Top-down programming
Petrography
CU = L
INDUCTION
Bottom-up programming
Facies
UNSUPERVISED
Principal Component Analysis
Cluster Analysis
SUPERVISED
Discriminant Function Analysis

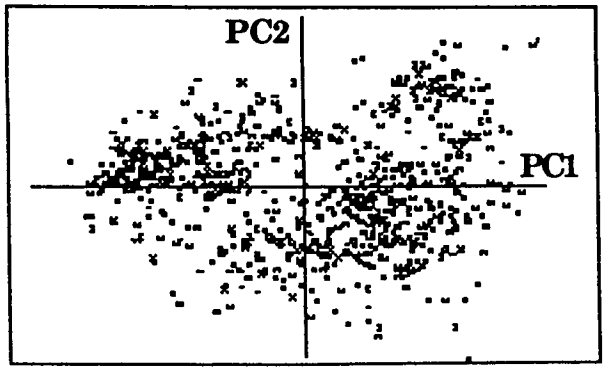
PRINCIPAL COMPONENT ANALYSIS EXAMPLE :
Pattern recognition of "electrofacies" through computation of the eigenvectors of standardized log variables of apparent grain density, apparent matrix volumetric cross-section, neutron porosity, thorium, uranium and potassium. The first two principal components account for 78% of the total variability, so that most of the six-variable information can be shown on a single crossplot. The principal component scores are themselves "logs" which may have diagnostic meaning as suggested by the eigenvector loadings.



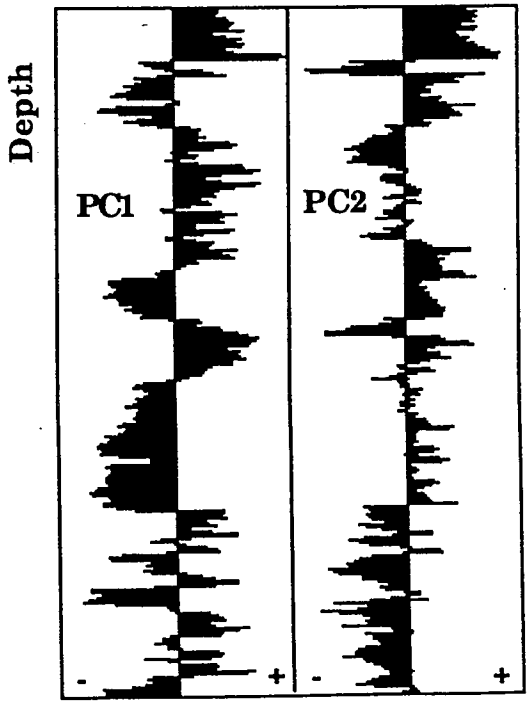
	PC1	PC2	PC3	PC4	PC5	PC6
RHOMAA	0.51	0.06	0.05	-0.16	-0.16	-0.83
UMAA	0.45	-0.05	-0.08	-0.79	0.04	0.41
CNL	0.39	0.36	0.72	0.25	0.32	0.20
Th	0.44	-0.30	0.00	0.38	-0.69	0.30
U	0.27	0.67	-0.63	0.25	0.07	0.12
K	0.35	-0.56	-0.28	0.29	0.62	-0.02
Eigenvalue%	59	19	8	7	5	2



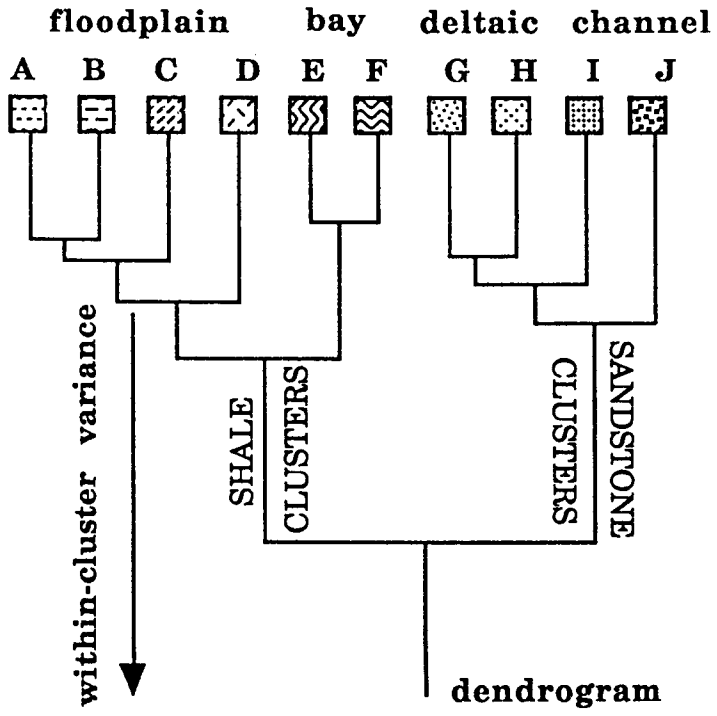
PC crossplot

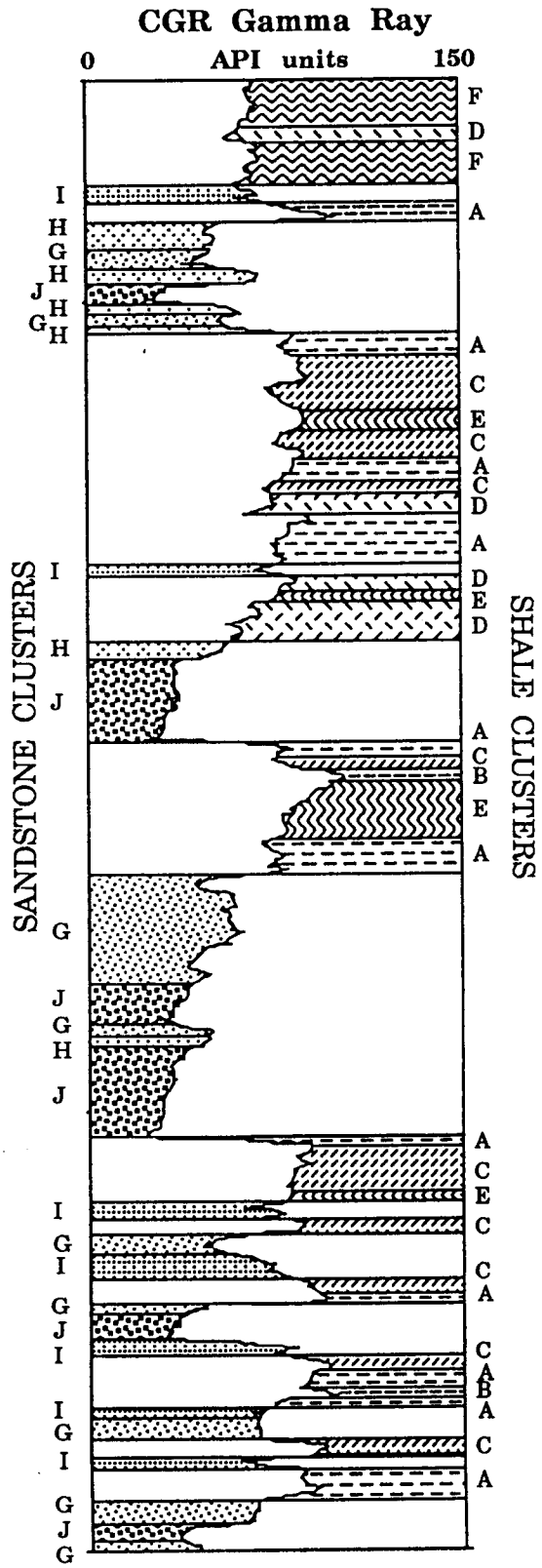


PC score logs

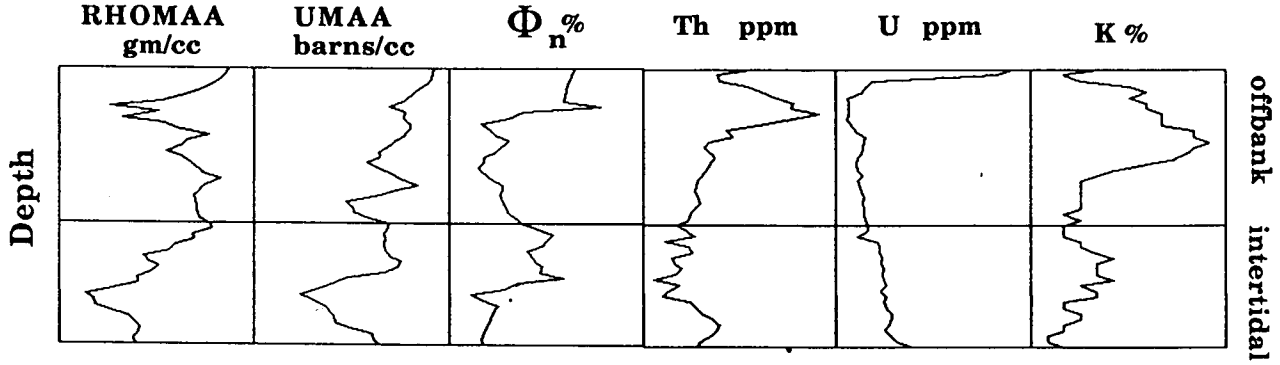


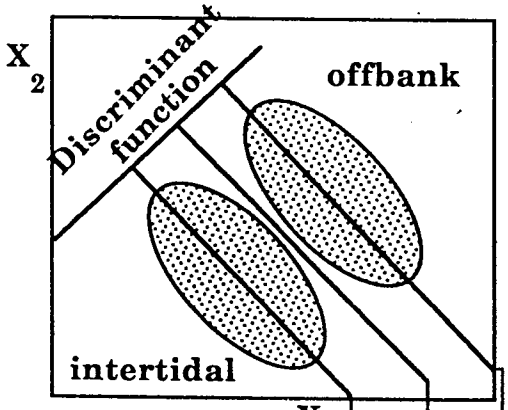
CLUSTER ANALYSIS EXAMPLE :
Location of clusters identified with different sedimentary environments in the Lower Cretaceous Dakota Formation.
The clusters are based on the log variables of apparent grain density, apparent matrix volumetric cross-section, neutron porosity, potassium, uranium and thorium.





DISCRIMINANT FUNCTION ANALYSIS EXAMPLE :
Discrimination between offbank and intertidal carbonate facies in the Winfield Limestone, based on six log variables of apparent grain density, apparent matrix volumetric cross section, neutron porosity, thorium, uranium and potassium.



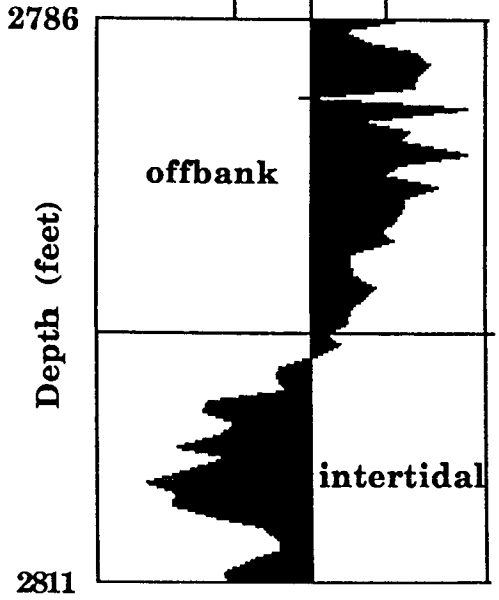


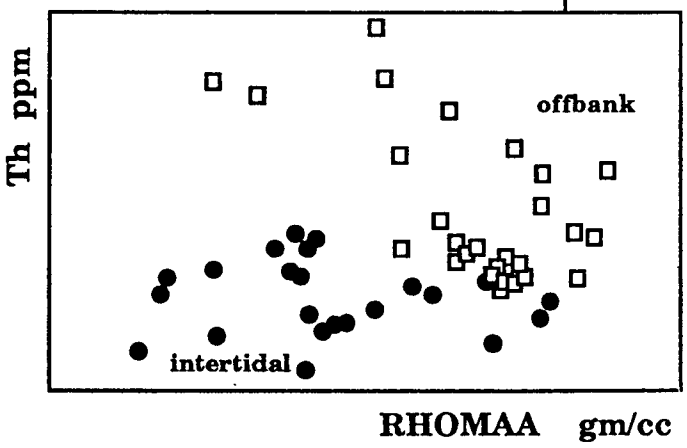
Discriminant function equation :

$Z = 131.9\ RHOMAA - 1.99\ UMAA + 0.33\ CNL\% + 5.9\ Th - 0.6\ U + 5.8\ K$

Relative contribution :

RHOMAA	48%
UMAA	10
CNL%	1
Th	28
U	2
K	12





BIBLIOGRAPHY OF LOG ANALYSIS STATISTICAL APPLICATIONS

Allen, J.R., 1979, Prediction of permeability from logs by multiple regression : Trans. SPWLA 6th. European Symp.

Berteig, V., Helgeland, J., Mohn, E., Langeland, T., and van der Wel, D., 1985, Lithofacies prediction from well data : Trans. SPWLA 25th Ann. Logging Symp., Paper TT, 25 pp.

Bolviken,E., Helgeland, J., Storvik,G., Siring, E., and van der Wel, D., 1988, Statistical permeability prediction from wireline logs by switching and Markov regime models : Trans. 11th. European Form. Eval. Symp., Paper Q.

Busch,J.M., Fortney, W.G., and Berry,L.N., 1985, Determination of lithology from well logs by statistical analysis : Preprint SPE 14301, 11 pp.

Doveton, J.H., 1986, Log Analysis of Subsurface Geology - Concepts and Computer Methods : John Wiley & Sons, 273 pp.

Doveton, J.H., and Bornemann, E., 1981, Log normalization by trend surface analysis : The Log Analyst, v. 22, no. 4, p. 3 - 8.

Etnyre, L.M., 1984, Practical applications of weighted least-squares methods to formation evaluation, Part I : The logarithmic transformation of non-linear data and selection of dependent variable : The Log Analyst, v. 25, no. 1, p. 11 - 21.

Etnyre, L.M., 1984, Practical applications of weighted least-squares methods to formation evaluation, Part II : Evaluating the uncertainty in least - squares results : The Log Analyst, v. 25, no. 3, p. 11 - 20.

Hempkins,W.B., 1977, Multivariate statistical approaches in formation evaluation : Trans. SPWLA 18th Ann. Logging Symp., Paper DD, 23 pp.

Heseldin, G.M., 1976, Discriminant analysis in petrophysics : Trans. SPWLA 17th Ann. Logging Symp., Paper OO, 10 pp.

Langeland,T., and Flotre,J.O., 1984, Transformation, interpolation, regression diagnostics, graphical displays, and Qv - porosity relations : Trans. SPWLA 25th Ann. Logging Symp., Paper KK, 21 pp.

Mendelson, J.D., and Toksoz, M.N., 1985, Source rock characterization using multivariate analysis of log data : Trans. SPWLA 26th Ann. Logging Symp., Paper UU, 21 pp.

Moss, B., 1987, Does principal components analysis have a role in the interpretation of petrophysical data? : Trans. SPWLA 28th Ann. Logging Symp., Paper TT, 25 pp.

Wendt, W.A., Sakurai,S., and Nelson, P.H., 1986, Permeability prediction from well logs using multiple regression in reservoir characterization, in Lake, L.W. and Carroll, H.B., Jr., eds., Reservoir Characterization : Academic Press, Orlando, p. 181 - 222.

Wolff,M., and Pelissier-Combescure, J., 1982, Faciolog - Automatic electrofacies determination : Trans. SPWLA 23rd Ann. Logging Symp., Paper FF, 23 pp.