SHORT COURSE on BASIC STATISTICS FOR PETROPHYSICISTS

John H. Doveton
Kansas Geological Survey
1996

Kansas Geological Survey
Open-file Report 96-60

Kansas Geological Survey
Open-file Report

Disclaimer

The Kansas Geological Survey does not guarantee this document to be free from errors or inaccuracies and disclaims any responsibility or liability for interpretations based on data used in the production of this document or decisions based thereon. This report is intended to make results of research available at the earliest possible date, but is not intended to constitute final or formal publication.

INTRODUCTION

SCALES OF MEASUREMENT

Scientific observations are related to four scales of measurement which are named *nominal, ordinal, interval* and *ratio* (listed in order of increasing information content).

Nominal Scale

Assignment to discrete categories which have no implicit ordering and no metrically defined boundaries; e.g., rock types, color.

Ordinal Scale

Discrete categorization with an inherent ordering; e.g., grade of show or porosity, stratigraphic age scale.

Interval Scale

Continuous or discrete numerical measurement in which distances between objects can be measured, but cannot be related to an absolute zero; e.g., spontaneous potential, structural elevation.

Ratio Scale

Continuous or discrete measurements with a definitive absolute zero; e.g., density, porosity, resistivity, permeability.

- Nominal and ordinal scales apply to non-metric discrete categorical data; interval and ratio scales are for metric discrete and continuous measurements.
- The greater information content of the higher grade scales extends the range of permissible statistics used to summarize the data and the precision of statistical inference based on them.

STATISTICS

Descriptive statistics

A variety of measures aimed at summarizing the characteristics of data sets (means, variances, correlations, etc.) together with pictorial representations of the data distributions (histograms, scatter plots, etc.).

Inferential statistics

The process of making generalizations or predictions concerning the phenomenon under study based on raw measurement variation and relationships between measured variables. Conclusions are drawn from limited information and used for making decisions under uncertainty. The logic is inductive, as inferences concerning the general are derived from a study of the observational particular.

All the values of interest (the universal set) are termed the *population*, for which summary measures are precise characterizations of the studied variables. These measures (mean, variance, etc.) are the *parameters* of the population.

It is usually only practical to measure a limited *sample* of the total population. Statistical measures of a sample are known as *sample statistics* and are *estimates* of the parameters of the parent population.

The sample must be representative of the total population in order for sample statistics to provide unbiased estimates of parameters. Random sampling provides a means by which every object in the population has an equal chance of being selected in the measured sample.

Parameters are conventionally denoted by *Greek* letters; sample estimates by *Roman*.

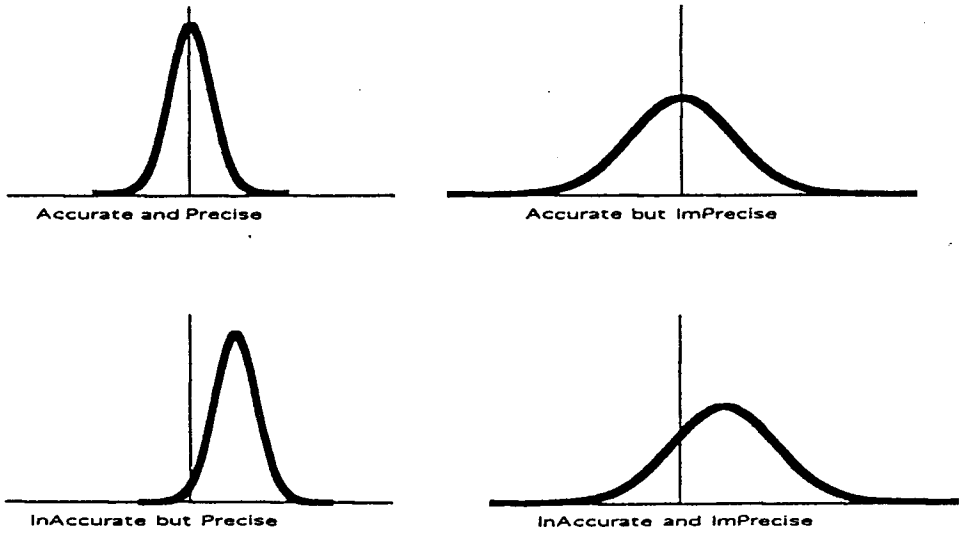
Univariate statistics are concerned with summarization and inferential analysis of a single variable measured on a sample of objects. *Multivariate statistics* marks an extension to several variables of measurement on each object and is the numerical description of variable interrelationships and inferences drawn from them.

The choice of descriptive and inferential methods is dictated largely by the scale of measurement of the observational variables and the geometric form of their distribution.

PRECISION AND ACCURACY

The terms *"accuracy"* and *"precision"* are sometimes (and wrongly) used interchangeably, but they have radically different meanings. The picture below should be helpful to clarify the distinction. Accuracy is a measure of how close to the true value the measured estimates tend to be. Precision is a measure of reproducibility or how well observations tend to repeat or cluster about some value. So it is possible to be extremely precise and dead wrong. Conversely, some methods can be quite accurate in the sense that on the average they tend to be right, but if the true answer has a narrow range, then the proportion of "misses" could be unacceptable.

Kimminau (1994) has a short but useful review of how accuracy and precision can be addressed when dealing with logs, cores, and reservoir estimations. He notes that while we obviously would like methods to be both accurate and precise, in practice, there is often a trade-off between the two. The two sources of error become compounded into the general term of *"uncertainty"*. Much of classical inferential statistics is directed at the analysis of error.



REFERENCE

Kimminau, S., 1994, Traceability - making decisions with uncertain data: The Log Analyst, v. 35, no. 5, p. 67-70.

PROBABILITY

The probability that an event will occur is registered on a scale ranging from zero (absolute impossibility) to one (absolute certainty). *A priori probabilities* can be set in advance of the occurrence of the event in situations where the physical constraints are exactly known (e.g. games of chance). *Empirical probabilities* are measured as a frequency ratio from an observed trial series where:

$$\text{probability of event} = \frac{\text{total number of occurrences of event}}{\text{total number of trials}}$$

For a finite trial series, the probability is a sample estimate, denoted by P. The population parameter of probability is symbolized by Π.

- If events A and B are possible outcomes in a trial series and cannot occur simultaneously, they are said to be *mutually exclusive*. Then the probability that either A or B will occur is the sum of their separate probabilities:

  $$\mathbf{P(A \text{ or } B) = P(A) + P(B)}$$

  This is the *additive rule of probability.*

- If events A and B are not mutually exclusive, but are independent of one another, then the joint probability that they will occur simultaneously is the product of their separate occurrence probabilities:

  $$\mathbf{P(A \text{ and } B) = P(A)*P(B)}$$

  This is the *multiplicative rule of probability.*

- When the occurrence of events A and B are dependent to some degree then their joint probability of occurrence is conditional and:

  $$\mathbf{P(A \text{ and } B) > \text{or} < P(A)*P(B)}$$

These concepts are reviewed and applied to the relationship between water saturations computed from log analysis and the results of drill-stem tests from carbonate formations in the Williston Basin (see WILLISTON...). Teti and Krug(1987) reported results for tests in 21 wells from the Mississippian Ratcliffe and Mission Canyon, the Devonian Nisku, Duperow, and Winnipegosis, the Silurian Interlake, and the Ordovician Gunton and Red River carbonates. They graded the test results as excellent show, very good show, good show, fair show, poor sow, no show. In this manual table, I have condensed the six outcomes to their categories of "commercial potential" (C) and "non-commercial" (N).

WILLISTON BASIN DST TEST FREQUENCIES THAT MATCH LOG ANALYSIS ESTIMATIONS OF WATER SATURATION (Sw) WITH FLUID RECOVERY OF COMMERCIAL (C) OR NON-COMMERCIAL (N)

Data from Teti and Krug (1987)

		C	N
Sw%	0-9		
	10-19	2	1
	20-29	4	3
	30-39	3	9
	40-49	5	10
	50-59	1	7
	60-69	1	11
	70-79		6
	80-89		2
	90-100		4

In a total of 69 tests, 16 had fluid recoveries that indicated commercial potential, while 53 test results were considered to be non-commercial. "Commercial" and "non-commercial" are mutually exclusive outcomes. Therefore the probability of a commercial result for a test in the Williston Basin, *based on these data* is:

$$P(C) = \frac{16}{69} = 0.23$$

and its complementary probability of a non-commercial result is:

$$P(N) = \frac{53}{69} = 0.77$$

The qualification of *based on these data* was inserted to point out that statistical probabilities are ideally computed from random samples. The exploration companies involved would clearly not want their tests to be random, but biased by indications from logs, cuttings, and mud shows to an expectation of successful outcome. If (say) the probability of a wildcat discovery in the Williston Basin is 0.1, then the computed P(C) suggests the prior indications have improved performance beyond random testing. Furthermore, the probabilities can be factored into economic analyses of risk and cost expectations.

The rows of the data table expand the information to water saturation as one of the log analysis variables that is considered prior to the decision on whether to run a test. The occurrence of any water saturation category and test result are not mutually exclusive. However, they can be either independent or dependent. If independent, then the log analysis of water saturation provides no additional information. If dependent, then we have the qualitative assurance that the log analysis computations are worthwhile. Perhaps, more importantly, the numbers give us a measure of performance and perhaps, a means to select a critical water saturation that matches the degree of risk that we are prepared to live with.

A condensed *contingency table* of outcomes is shown below that relates test results to whether the log estimate of water saturation was greater or less than 50%:

	C	N	Row totals
Sw<50%	14	23	37
Sw>50%	2	30	32
Column totals	16	53	69

The joint probability of a log analysis estimate of Sw<50% and a commercial test result is the joint frequency (14) divided by the grand total of tests (69):

$$P(Sw < 50\% and C) = \frac{14}{69} = 0.20$$

What would be the expected probability if they were independent? The independent expectation can be calculated from the multiplicative rule of probability:

$$P(Sw<50\% \text{ and } C) = P(Sw<50\%) * P(C)$$

The *marginal probability* of Sw<50% is:

$$P(Sw < 50\%) = \frac{37}{69} = 0.54$$

The marginal probability of commercial potential is:

$$P(C) = \frac{16}{69} = 0.23$$

Therefore, their *unconditional joint probability* is:

$$P(Sw < 50\% and C) = 0.54 * 0.23 = 0.12$$

The observed joint probability is nearly double this expectation, so that there is a conditional relationship between this log estimate and the test result.

We can also talk in terms of *conditional probability* or what is the probability of event A **given** that event B has already occurred? This expression would be symbolized as P(A/B). This concept has immediate application to our example. How does the use a 50% water saturation cutoff to screen tests compare with tests in general? From the contingency table:

$$P(C / Sw < 50\%) = \frac{14}{37} = 0.38$$

which is an improvement on the unconditional probability of:

$$P(C) = \frac{16}{69} = 0.23$$

If we examine the contingency frequencies (see WILLISTON...) and the associated probabilities carefully, we learn some interesting conclusions. For example, a 50% water saturation cut-off will ensure that very few oil or gas zones will go untested (2 out of 69). However, more tests will be non-commercial (23) than those that have commercial potential (14). But if we make the water saturation cutoff more stringent, then the success rate will go up, but we will fail to test many commercial zones.

Over and beyond the table, there are other considerations. What are the implications if the exploration company has a much greater chance of locating oil reservoirs than the average in their use of predrill information (i.e. P(C)>>0.23)? How would this compare with a company that drills only random (not by choice) wildcats (probably P(C)<0.23)? *Bayesian probability* has been used widely to address this kind of question.

The Reverend Thomas Bayes proposed *Bayes' theorem* which lead to the relationship:

$$P(B_i / A) = \frac{P(A / B_i)P(B_i)}{\sum P(A / B_i)P(B_i)}$$

The equation gives a way of estimating the conditional probability P(B/A) when we only know P(A/B). In this example, we would like to know the probability of a commercial result given a calculated water saturation. The data table gives us the necessary information.

The Bayesian equation can be made more specific as:

$$P(C / Sw) = \frac{P(Sw / C)P(C)}{P(Sw / C)P(C) + P(Sw / N)P(N)}$$

Let us suppose that the industry wildcat rate of penetrating a commercial zone is 0.1 and we wish to evaluate the performance of using a 50% water saturation cut-off. Then, P(C)=0.1 and:

$$P(C/Sw<50\%) = \frac{(14/16)*(0.1)}{(14/16)*(0.1) + (23/53)*(0.9)} = 0.18$$

Notice that this is even worse than our original estimate of a marginal probability of P(C) of 0.23 (in other words using no Sw information at all) ! But that is because we stepped back to a broader situation of a discovery rate of 10%, which would mean that 90% of what we drilled was non-commercial and so we can expect a high number of non-commercial tests.

A situation where P(C) was actually 0.23 could be proposed where obviously poor zones had been eliminated using a variety of criteria and we wish to find the improvement (if any) that would result from using the cutoff value of 50% water saturation. Then the Bayesian prediction is:

$$P(C/Sw<50\%) = \frac{(14/16)*(0.23)}{(14/16)*(0.23) + (23/53)*(0.77)} = 0.38$$

which is the same number that was computed by classical probability earlier, and does show a systematic improvement.

At yet another extreme, notice that if prospects were drilled at a 100% success rate (P(C)=1.00), the cutoff would provide a perfect record of commercial tests on every call. The log analyst for this company would look a lot sharper than the one who works for one whose prospects were duds, even though the two analysts might be using the same cutoff. Notice also that the log analyst of the better company would have a little secret -- by using the 50% cutoff, two out of every 16 wells would be untested and written off even though they were commercial.

In these calculations, we have taken a specific water saturation cutoff of 50% as an example. The consequences of other cutoffs can be computed from the table by substituting the appropriate frequencies in the equations that have been described. So, for example, a Bayesian analysis of alternative cutoffs and with a marginal commercial probability of 0.1 gives the following results:

Sw<20%	P(C/Sw) = 0.42
Sw<30%	P(C/Sw) = 0.36
Sw<40%	P(C/Sw) = 0.20

Sw<50%	P(C/Sw) = 0.18
Sw<60%	P(C/Sw) = 0.16
Sw<70%	P(C/Sw) = 0.13
Sw<80%	P(C/Sw) = 0.11
Sw<90%	P(C/Sw) = 0.11
Sw<100%	P(C/Sw) = 0.10

The choice of cutoff is obviously dictated by the desire to maximize the number of successes while minimizing the number of failures. Ultimately, all the contingencies must be weighted according to their costs and this can be done in dollar amounts to a certain degree, although measures of "utility" are probably more realistic.

REFERENCES

Teti, M.J., and Krug, J.A., 1987, Log analysis methods and empirical derivation of net pay parameters for carbonate reservoirs in the eastern Montana part of the Williston Basin: The Log Analyst, v. 28, no. 3, p. 259-281.

PICAROON SANDSTONE

PICAROON SANDSTONE PREAMBLE

The Picaroon field is located on the Corsair trend of offshore Texas and has estimated gas reserves of 400 bcf (Vogler and Robison, 1987). The structure of the field is an anticline with a NE-SW axis, that is extensively faulted and truncated by an unconformity to the east. The reservoir sandstones are part of a middle and upper Miocene regressive sequence. They are feldspathic to sublithic sandstones and are considered to represent deltaic inner fringe deposits (Taylor, 1990). Calcite cement and clay content has reduced porosities and permeabilities of many sandstones on the Corsair trend. However, diagenetic dissolution of calcite in sandstones of the Picaroon field has resulted in greatly improved reservoir quality.

Taylor (1990) reported thin-section point counts of composition together with porosity and permeability for Picaroon field sandstones. The compositional variables recorded were: quartz, lithic fragments, detrital calcite (limestone rock and minor marine shell fragments), clay, quartz cement, calcite cement, ankerite cement. Because this is a demonstration data set in a basic statistics course manual, the components have been consolidated into a smaller number of variables, but which still retain the character of the original data. Lithic fragments and detrital calcite have been combined into a single "LITH" component; quartz, calcite, and ankerite cements have been aggregated into a single "CEMENT". The composition of 39 Picaroon sandstones are tabulated in terms of quartz, lith, clay, cement, and porosity, and are listed with sample depth and permeability (see PICAROON SANDSTONE DATA TABLE).

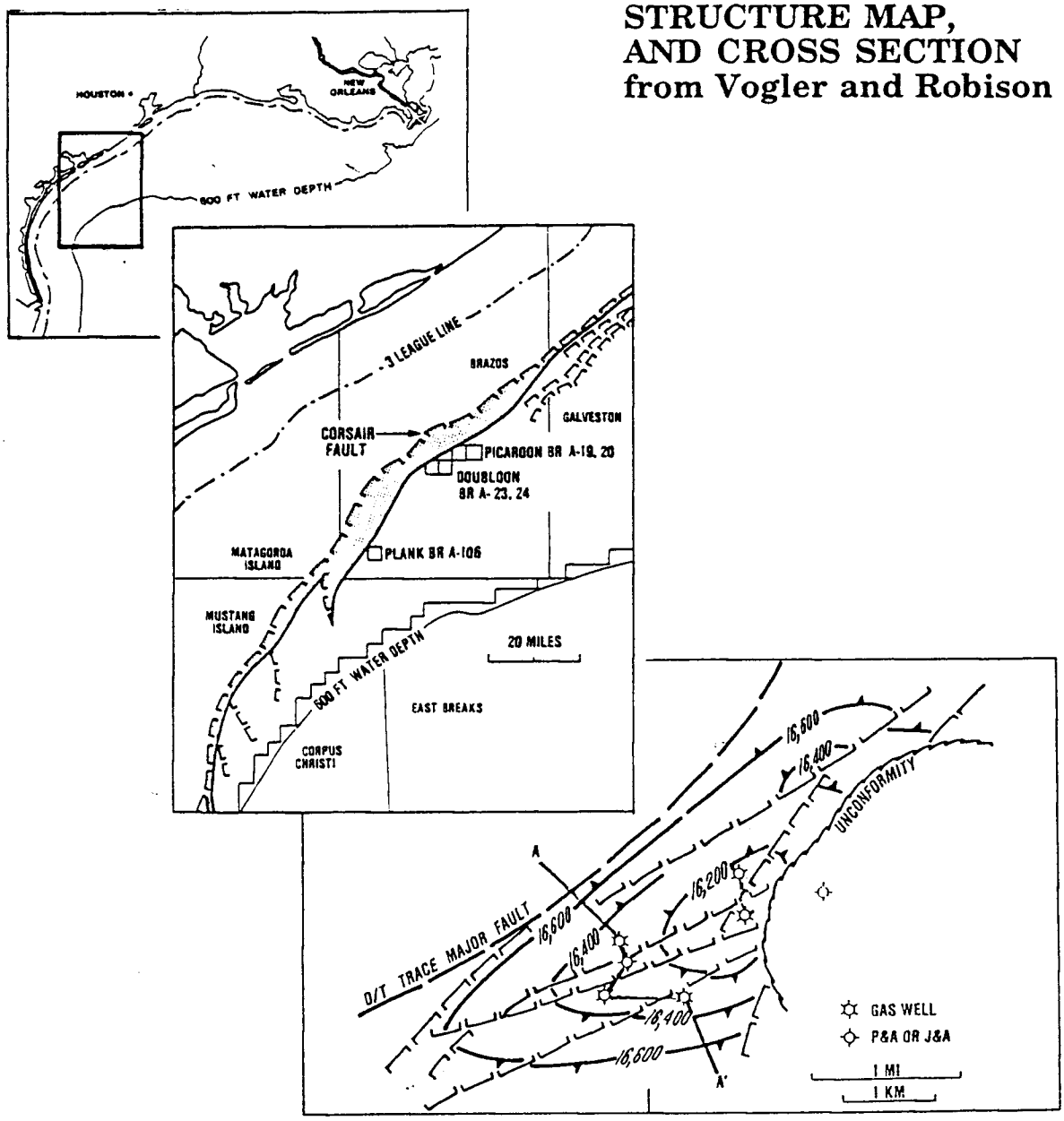
The depth range of the 39 sandstones show distinctive groupings at four levels (see PICAROON SANDSTONE SAMPLE DEPTH LOCATIONS). Four informal subgroups have been named as "Unit 0" (Sample #1), "Unit 1" (samples #2 to #21), "Unit 2" (samples #22 to 29), and "Unit 3" (samples #30 to #39). In the following demonstrations of statistical techniques, the Picaroon sandstones will be analyzed both in aggregate and as subdivisions.

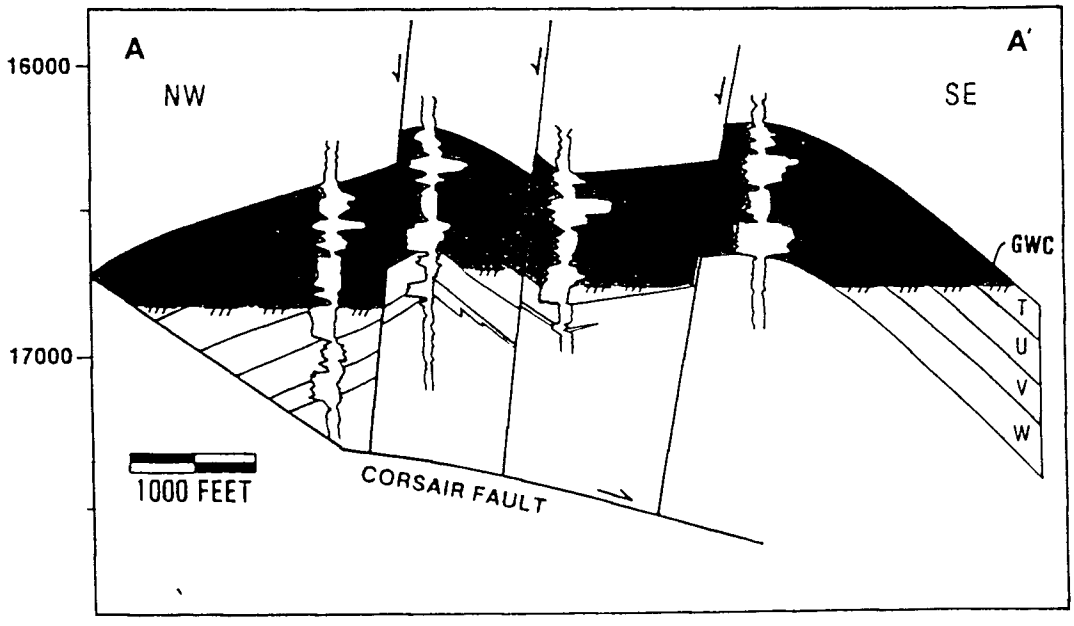
REFERENCES

Taylor, T.R., 1990, The influence of calcite dissolution on reservoir porosity in Miocene sandstones, Picaroon field, offshore Texas Gulf Coast: Jour. Sed. Petr., v. 60, no. 3, p. 322-334.

Vogler, H.A., and Robison, B.A., 1987, Exploration for deep geopressured gas: Corsair Trend, offshore Texas: AAPG Bulletin, v.71, no. 7, p.777-787.

PICAROON FIELD LOCATION, STRUCTURE MAP, AND CROSS SECTION
from Vogler and Robison (1987)





PICAROON FIELD SANDSTONE THIN-SECTIONS
from Taylor (1990)

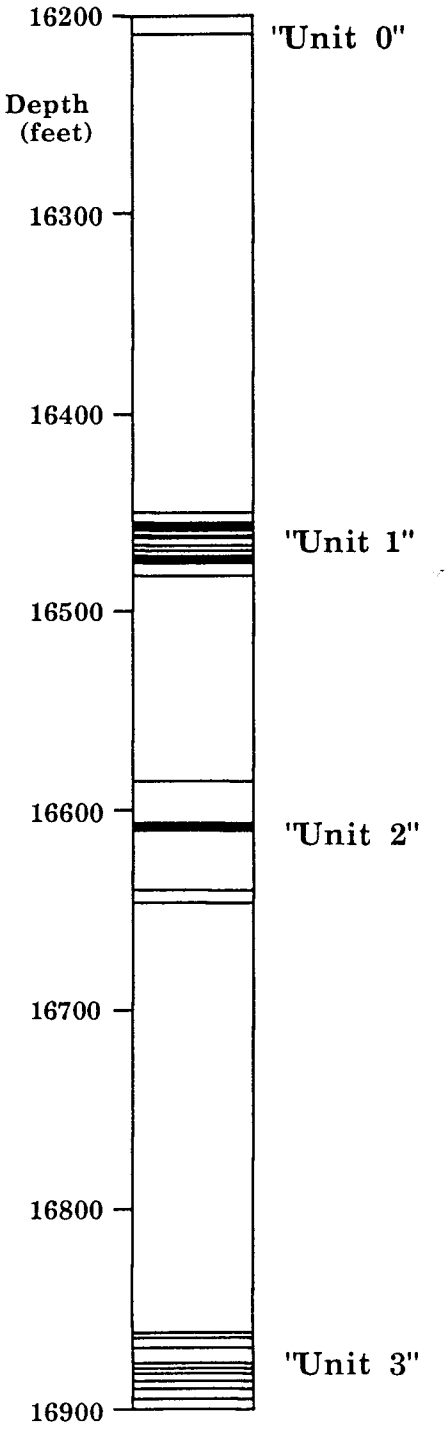
A. A low-porosity sandstone with approximately 30% calcite cement.

B. A sandstone (same bed as A.) with approximately 27% porosity and little calcite cement. Note that the volume of porosity in B. is approximately equal to the volume of cement in A.

PICAROON SANDSTONE DATA TABLE

ID#	Depth (feet)	Porosity %	Permeability md	Quartz %	Lith %	Clay %	Cement %
1	16209.0	19.6	43.00	39.2	19.3	4.5	17.4
2	16450.0	23.3	525.00	43.4	21.9	3.1	8.4
3	16454.5	13.8	0.86	38.2	20.6	1.7	25.7
4	16455.0	18.2	4.90	41.1	23.1	0.6	17.0
5	16455.5	17.6	2.60	34.4	23.7	1.0	23.3
6	16456.0	19.6	2.90	46.3	15.5	1.5	17.0
7	16457.0	19.8	3.80	44.9	20.2	0.3	14.8
8	16457.5	21.2	5.40	36.2	26.6	1.6	14.4
9	16458.0	22.0	87.00	40.8	25.5	0.3	11.4
10	16459.0	21.4	73.00	40.2	22.2	2.4	13.8
11	16459.0	21.4	73.00	38.5	25.7	0.6	13.8
12	16461.0	18.2	1.80	42.0	25.4	3.6	10.7
13	16462.0	20.0	5.30	40.2	25.9	3.8	10.1
14	16463.0	20.4	6.90	44.4	19.8	2.6	12.8
15	16466.5	12.5	4.60	36.5	21.4	6.2	23.4
16	16469.0	15.0	21.00	34.1	22.2	6.4	22.2
17	16472.0	16.8	7.80	33.7	25.3	1.7	22.5
18	16473.0	17.2	7.20	35.8	22.8	3.4	20.8
19	16474.0	19.3	17.00	31.9	25.3	2.5	21.0
20	16475.0	11.1	1.00	37.4	25.1	2.0	24.4
21	16482.0	14.1	5.30	34.4	21.8	1.6	28.0
22	16586.0	10.0	0.01	38.6	19.1	5.8	26.5
23	16606.0	19.5	71.00	38.0	20.9	4.2	17.3
24	16607.0	20.0	82.00	37.3	22.3	6.0	14.4
25	16607.5	11.5	0.15	37.2	24.5	2.8	24.0
26	16609.0	13.7	2.40	37.7	28.9	3.3	16.4
27	16610.5	9.0	0.08	32.1	27.1	5.4	26.4
28	16640.0	3.5	0.06	38.5	18.4	4.0	35.6
29	16646.0	20.3	0.13	38.8	20.1	4.7	16.1
30	16860.9	2.8	0.03	44.9	15.0	5.2	32.1
31	16863.3	2.0	0.03	47.1	22.1	0.6	28.2
32	16868.4	11.2	2.10	51.6	20.6	0.3	16.2
33	16869.0	21.1	215.00	50.4	12.2	3.3	12.9
34	16876.0	28.2	896.00	43.8	22.2	1.6	4.1
35	16879.0	27.8	1150.00	50.2	15.5	2.5	4.1
36	16881.0	27.4	1210.00	48.1	19.0	1.3	4.2
37	16886.0	27.2	1100.00	49.8	17.2	2.4	3.4
38	16890.0	25.7	399.00	44.6	23.5	1.9	4.3
39	16895.0	27.1	838.00	49.9	18.1	1.6	3.3

PICAROON SANDSTONE SAMPLE DEPTH LOCATIONS



PICAROON SANDSTONE POROSITY HISTOGRAMS

A plot of the Picaroon sandstone porosities versus depth shows the vertical variability of porosity within the reservoir profile as penetrated by the well (see PICAROON SANDSTONE POROSITY PROFILE). A variety of graphical techniques can be used to summarize the overall distribution of values which are difficult to pick out from a data tabulation. So, for example, a density stripe plot of porosities shows immediately the location of all observations on their measurement scale (see BASIC PRESENTATIONS...). Although density stripe plots work well for small and moderate sample sizes, there is an increased tendency for stripe overlap and formation of black bars with larger samples. Histograms are used more commonly used as a means to show the relative concentration or density of data values along their measurement scale.

Bar charts are widely used to summarize frequencies of different discrete categories. Although histograms look the same, they are used to show sample density of continuous variables. They are useful in characterizing overall distribution shape and locating a mode (or modes) in the data. However, unlike bar charts, the shape of the histogram will be controlled not only by data variability, but also the user's (or software's!) choice of bin width and bin origin. The bin width specifies the incremental scale range of data to be counted for each histogram bar. The bin origin marks the boundary value of the lowest bin on the scale. Successive bin boundaries will be located at multiples of the bin width relative to the bin origin.

Too coarse a bin width causes an oversmoothing of the data; too narrow a bin results in poor generalization of the data density. This can be seen from a comparison of alternative histograms drawn for the Picaroon sandstone porosities (see ALTERNATIVE HISTOGRAMS...). A number of rules have been devised to select appropriate bin width. The most widely known is Sturges' rule (Sturges, 1926), where the number of bins, k, is given by:

$$k = 1 + \log_2 n$$

where n is the number of observations. In the case of the Picaroon sandstones, the number of observations (n) is 39, so k= 6 (taken to the nearest integer). The minimum porosity is 2.0%, the maximum is 28.2% which makes an effective range of thirty porosity units with bin origin at zero. Therefore, Sturges' rule suggests a histogram of six bins of bin-width 5% to span the porosity range. Sturges' rule is based on a binomial model for normally distributed data. When data are skewed, the rule tends to underestimate the number of bins. Some software packages apply Sturges' Rule as a default; others apply a pragmatic estimation of: $k = \sqrt{n}$ which gives greater number of bins at large sample sizes.

The edges of the bins mark abrupt discontinuities in the subdivision frequency counts of what is a continuous variable. Changes in boundary locations may adversely affect the histogram display, so that features may be less real than they appear. Some of these artifacts can be seen in the comparison of the histograms with the stripe density plot. Notice, for instance, how the form of the central mode changes when a bin boundary coincides with 20% porosity. The cluster of observations around this value results in a subdivision into two adjacent bins and a flatter shape to the overall histogram. This problem of boundary selection is well-known and authors such as Scott (1992) suggest the use of an "average shifted histogram" (ASH) to overcome the discontinuity effect.

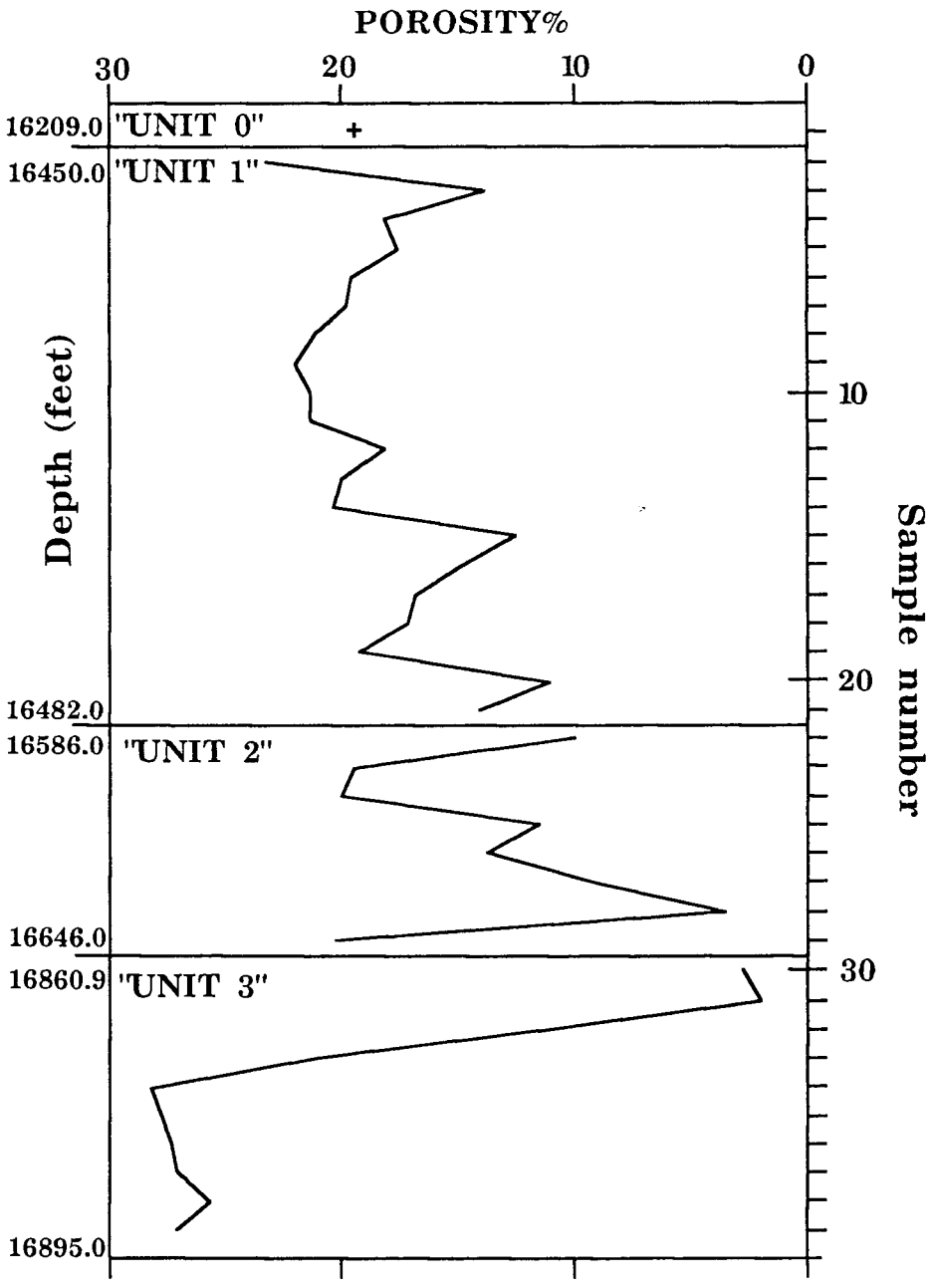
In summary, histograms are a widely used and useful graphic summary of the density distribution of data. Ultimately, they are crude density estimators and should be checked carefully with alternative bin widths and origins before pronouncements are made concerning systematic modes or other shape features.

REFERENCES

Scott, D.W., 1992, Multivariate Density Estimation: Theory, Practice, and Visualization: John Wiley, New York, 317 pp.

Sturges, H.A., 1926, The choice of a class interval: J. Amer. Statist. Assoc., v. 21, p. 65-66.

PICAROON SANDSTONE POROSITY PROFILE

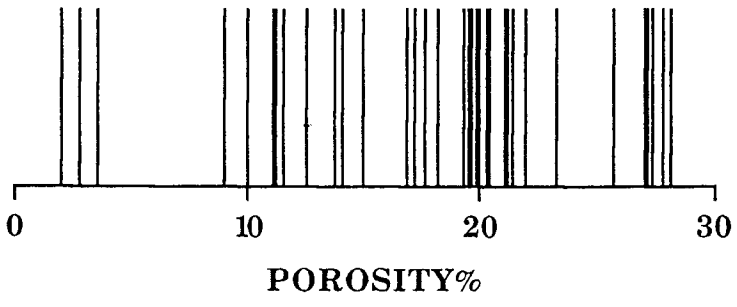


BASIC PRESENTATIONS OF PICAROON SANDSTONE POROSITY DATA

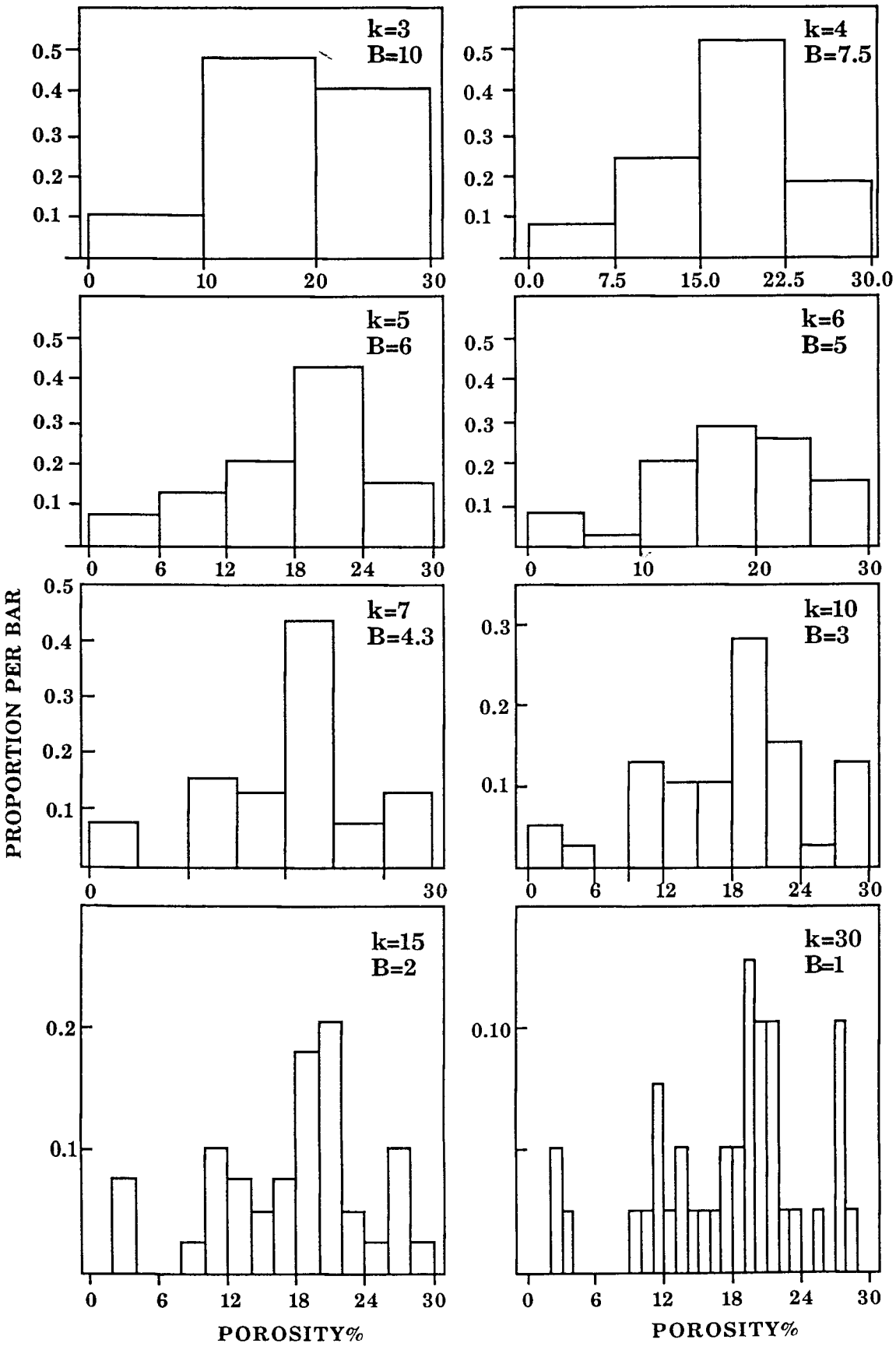
RAW DIGITS

#	Φ	#	Φ	#	Φ
1	19.6	14	20.4	27	9.0
2	23.3	15	12.5	28	3.5
3	13.8	16	15.0	29	20.3
4	18.2	17	16.8	30	2.8
5	17.6	18	17.2	31	2.0
6	19.6	19	19.3	32	11.2
7	19.8	20	11.1	33	21.1
8	21.2	21	14.1	34	28.2
9	22.0	22	10.0	35	27.8
10	21.4	23	19.5	36	27.4
11	21.4	24	20.0	37	27.2
12	18.2	25	11.5	38	25.7
13	20.0	26	13.7	39	27.1

DENSITY STRIPE PLOT



ALTERNATIVE PICAROON SANDSTONE POROSITY HISTOGRAMS SET BY NUMBER OF BINS (k) AND MATCHING BIN-WIDTHS (B)



PICAROON SANDSTONE POROSITY QUANTILE PLOT

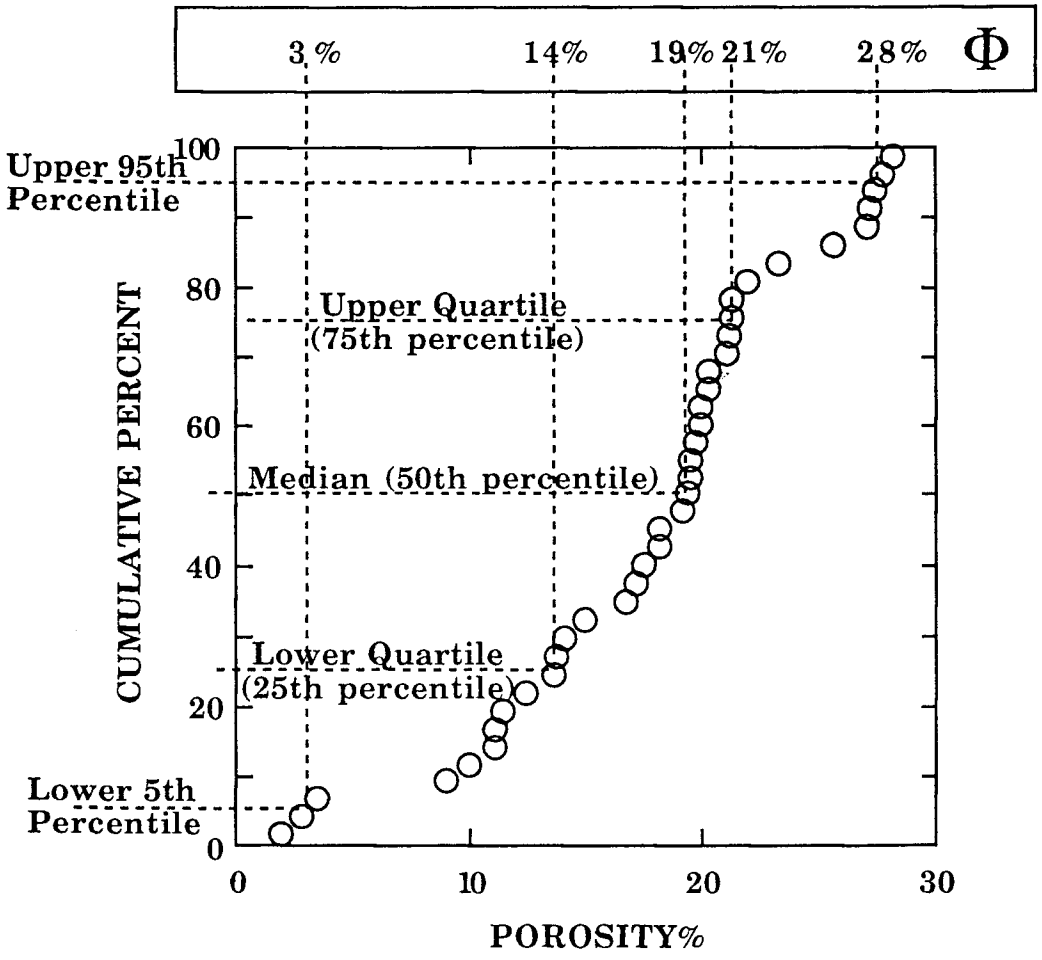
A quantile plot (abbreviated as Q-plot and also known as a cumulative plot) shows the individual observations plotted against their cumulative occurrence. A "quantile" of a sample is the value of the observation that matches a given proportion of the sample. A Q-plot of the Picaroon sandstone porosities (see QUANTILE PLOT...) can be used both to characterize the form of the porosity distribution and to generate useful summary statistics. The fundamental quantile is the median which corresponds to the 50th percentile. The median Picaroon sandstone porosity is 19%. Based on this sample, it is estimated that 50% of porosities should be greater than this value; 50% should be less. The median is a measure of central tendency that can be used as the most basic descriptor of a distribution of observations. (Other measures of central tendency are the mean and the mode, whose properties will be discussed later.)

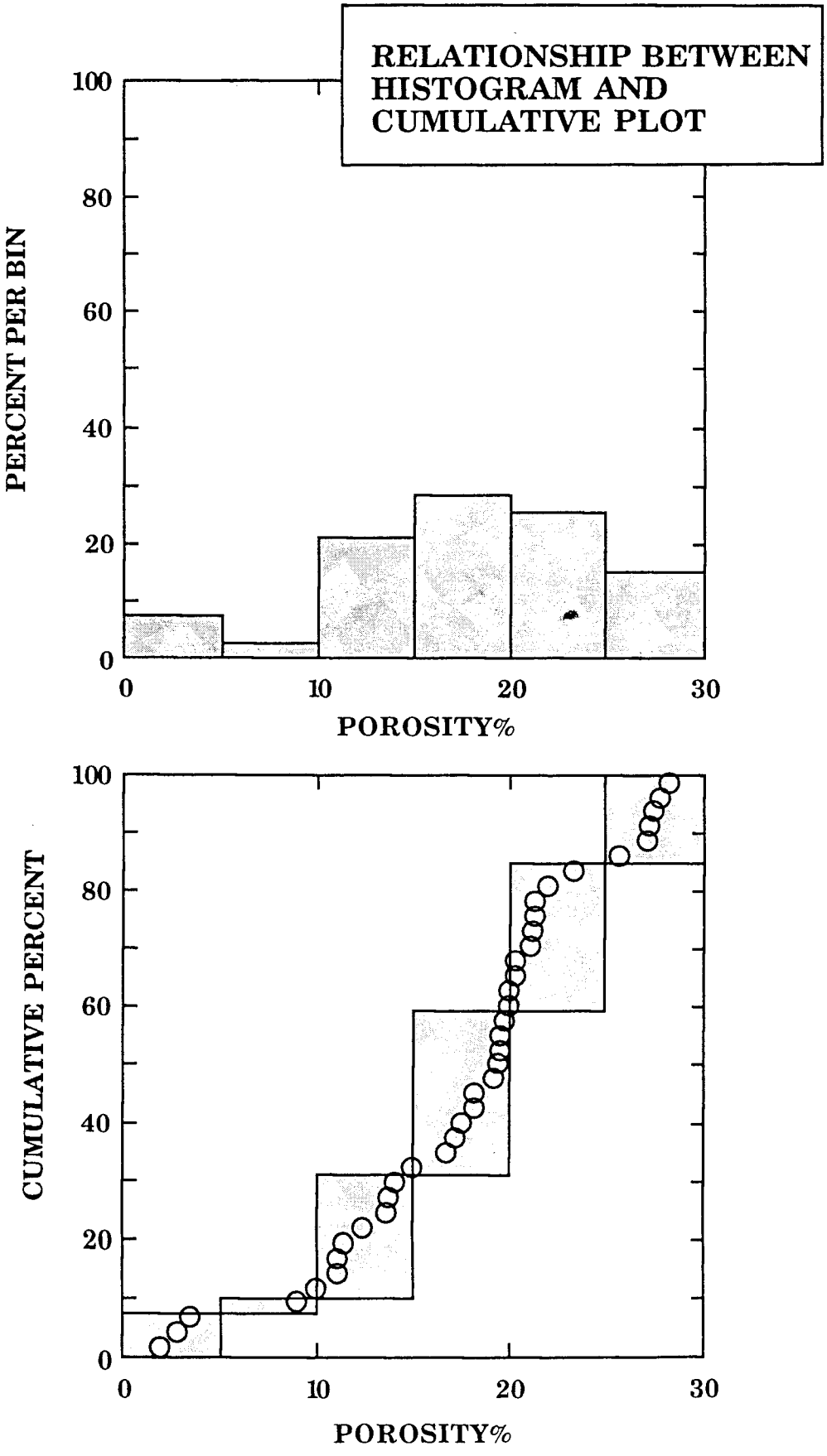
The lower quartile (25% of the sample) is a porosity of 14% and the upper quartile (75% of the sample) is 21%. The two quartiles can be thought of as the "medians" of the two sample halves split by the median. Their values give a general idea of the spread of the data. Finally, low and high percentiles (such as the lower 5th and upper 95th percentiles) indicate the observation values in the tails of the distribution. Although these values are usually close to the range (the lowest and highest values), they are much more stable when making comparisons between different samples. The maximum and minimum values will reflect any rogue or outlier observations.

There is a direct relationship between quantile plots and histograms (see RELATIONSHIP...); they are simply alternative ways of graphing the data. Note that any of the alternative histograms discussed earlier could have been generated by the single Q-plot. The overall shape and occurrence of modes is usually easier to see on histograms (provided good choices of bin width and boundaries are made as discussed earlier). The same information is present on the Q-plot, but takes a more practiced eye to make them out. A normal distribution (or any distribution with a central mode and extended tails) will plot as an ogive or S-shape on the quantile plot. A uniform distribution will generate a straight line of values. Later in the manual we will examine probability (or P-plots) where the cumulative percent axis is scaled to conform with the expectations for a given distribution (usually the normal).

The Q-plot Picaroon sandstone porosities show a generalized ogive shape that contrasts the bulk of the central observations with the less frequent observations in the tails. Note the extended and particularly steep trend at about 20% porosity that marks the high proportion of observations close to this value. The shapes of the tails may also be significant. At the lower end, the clump of tight sandstones have pore spaces that are almost completely occluded with cement. The steep trend may reflect the fact that the scale is truncated at zero; negative porosities are obviously impossible. At the higher end, the steep break at about 28% porosity may reflect the approach to some kind of natural limit -- an expectation of porosity for cement-free Picaroon sandstones.

QUANTILE PLOT (Q-Plot or cumulative plot) OF PICAROON SANDSTONE POROSITY




RELATIONSHIP BETWEEN
HISTOGRAM AND
CUMULATIVE PLOT
PERCENT PER BIN
100
80
60
40
20
0
0
10
20
30
POROSITY%
CUMULATIVE PERCENT
100
80
60
40
20
0
0
10
20
30
POROSITY%

PICAROON SANDSTONE POROSITY BOX PLOTS

The method of box plots were introduced by Tukey (1977) as a simple graphical means to show the distribution form of a single variable. There are a variety of conventions on how box plots are drawn, but the one used to illustrate Picaroon sandstone porosity is fairly typical (see BOX PLOTS...). The box features are drawn with reference to a measurement scale of the variable. The limits of the box are called "hinges" and are matched with the upper and lower quartiles. The width of the box is therefore the interquartile range, often known as the "spread". The median is marked by a vertical line within the box. Axes that extend beyond the box are called "whiskers" (hence the alternative name of "box-and-whisker plot") with limits at specified low and high percentile extremes. In this example, the lower 5th and upper 95th percentile have been used to set the ends of the whiskers. Extreme observations that constitute outliers are discriminated as values that lie beyond a "fence" value. Their locations are usually shown by individual symbols (stars, crosses etc). The fence is set at a multiple of spread distances (commonly 1.5 or 3) above and below the box hinges.

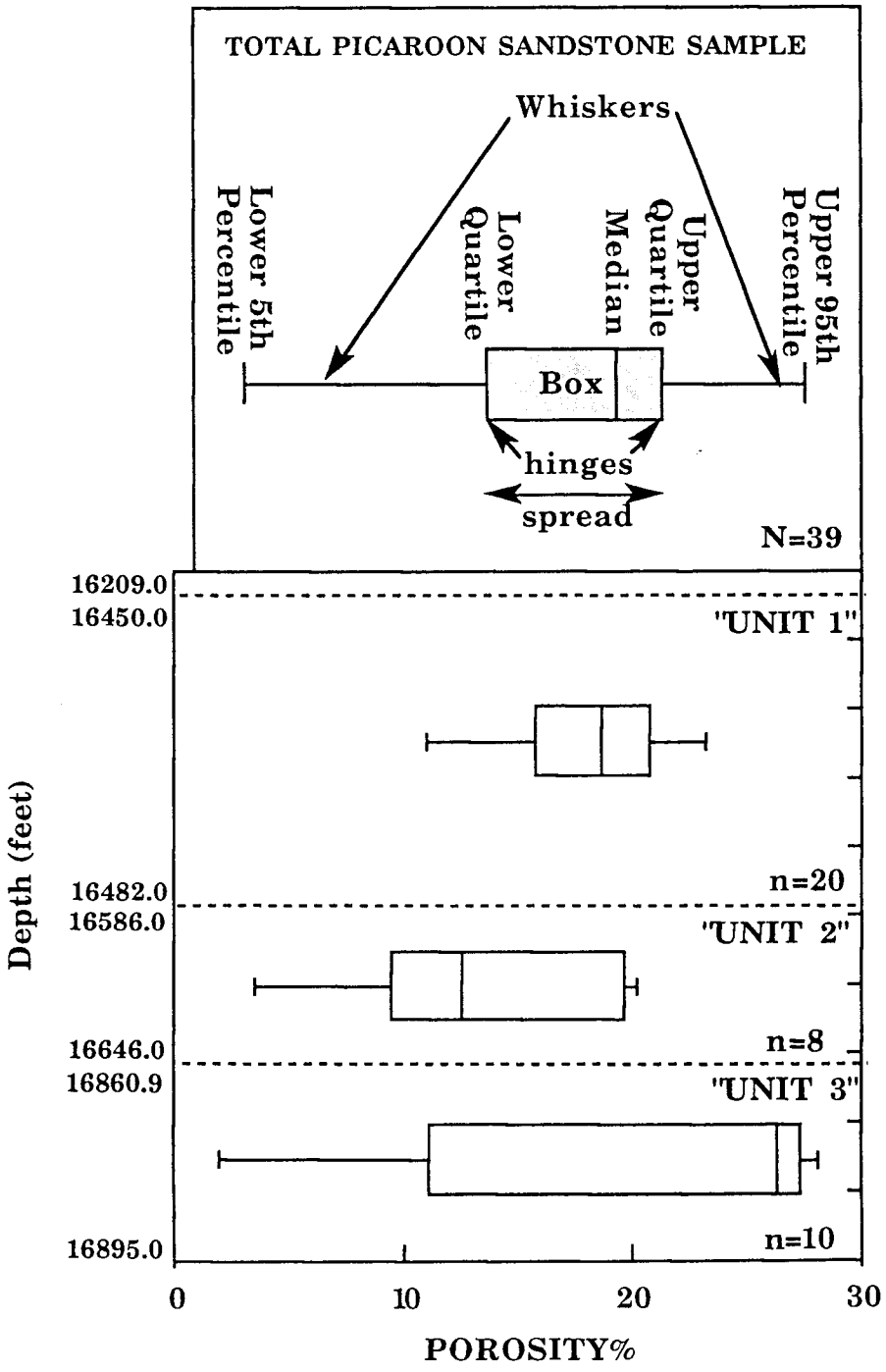
Box plots are shown for porosity distributions within the unit subdivisions of the Picaroon sandstone. Their graphic summaries highlight both their general distribution character as well as similarities and differences between them in a succinct manner. The box of "Unit 1" porosities show a compact, fairly symmetrical distribution centered on a porosity of about 18%. The boxes of "Units 2 and 3" are markedly different with strongly asymmetric (skewed) characters.

Notice that the box plot is a graphic expression of the quantile parameters from a quantile plot (see RELATIONSHIP...). The use of the box plot allows quick assessments and comparisons to be made of the quantile descriptors of Q-plots for moderate or large numbers of samples.

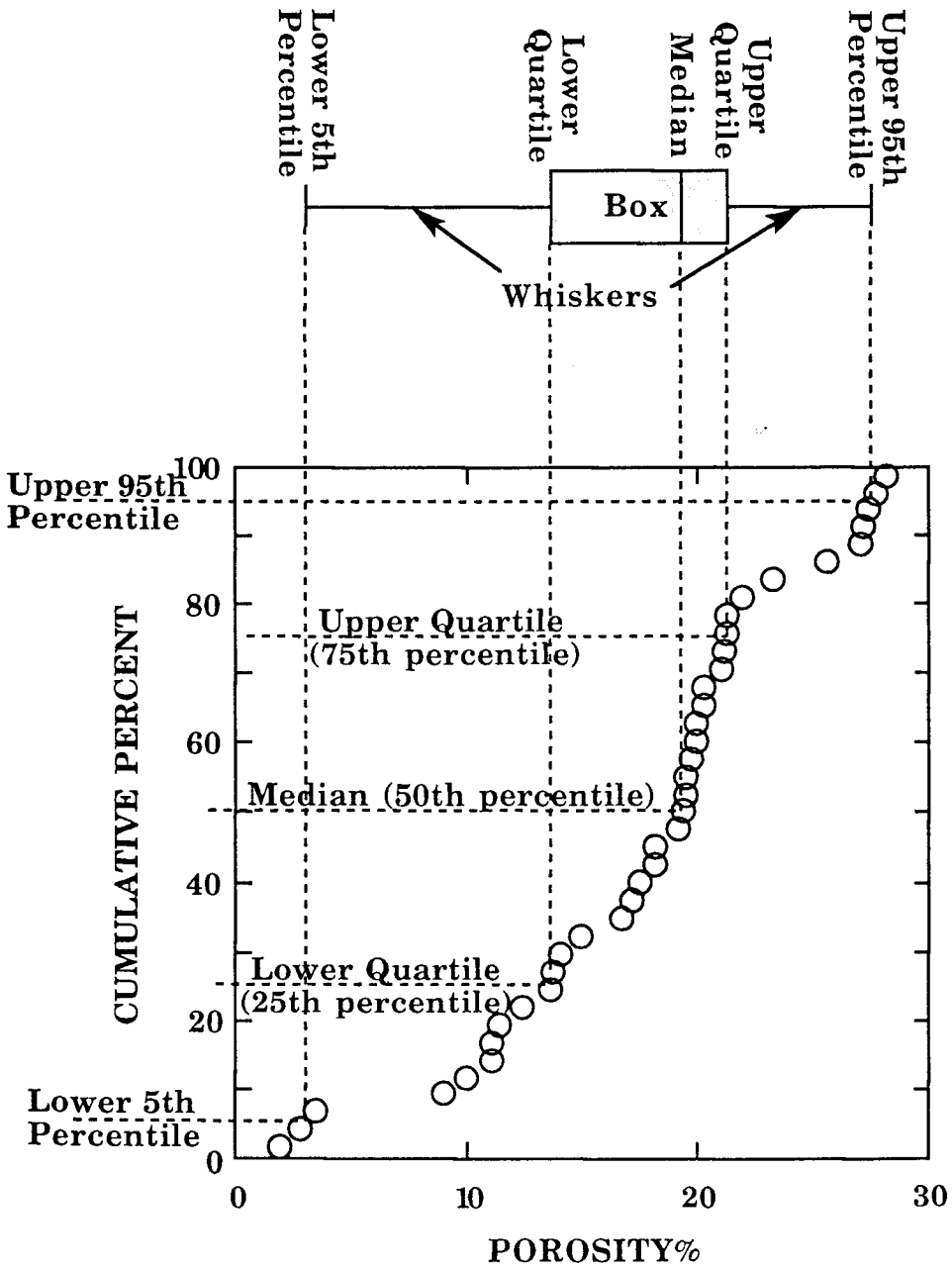
REFERENCE

Tukey, J.W., 1977, Exploratory Data Analysis: Addison-Wesley, Reading, Mass.

BOX PLOTS OF PICAROON SANDSTONE AND UNIT SUBDIVISIONS



RELATIONSHIP BETWEEN BOX PLOT AND QUANTILE PLOT



DESCRIPTORS OF LOCATION (measures of central tendency)

The most basic parameter of a distribution is a statistic that expresses a representative value and is a measure of distribution centrality. Three alternative measures are commonly used: the mode, the median, and the mean. Each has useful properties that make it appropriate in different applications.

The mode (Mo)
The mode is the most frequently occurring value in the distribution. It is applicable to all four measurement scales. It is the only central measure available for nominal data. The ideal of a mode is that it represents the most *typical* value. However, a modal class may have marginally higher frequency than other classes. Alternatively, a distribution may have several modes. Consequently, the mode may not be a very stable estimate of centrality.

The median (Md)
The median is the value that subdivides the distribution into two equal halves. Fifty percent of the observations have value higher than the median; fifty percent have lower values. The median is a stable estimate of centrality and can be preferable to the mean when the distribution has extreme outliers.

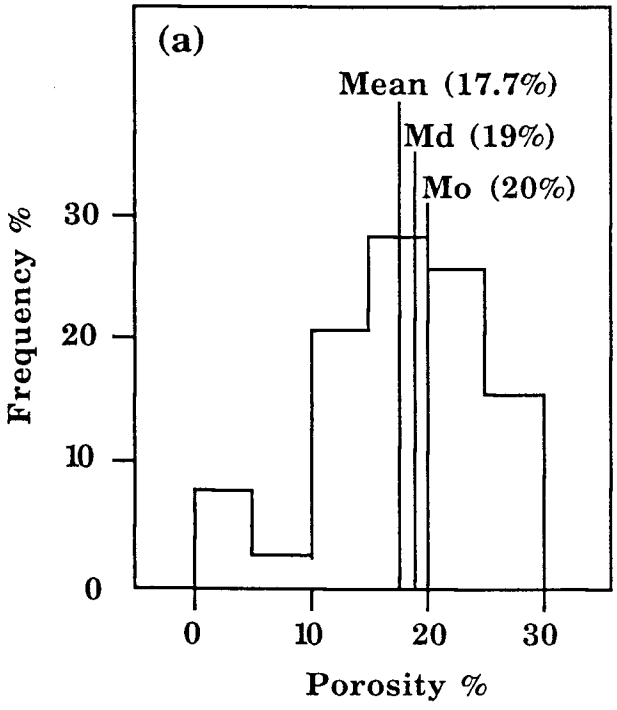
The mean
The mean of a sample is the arithmetic average, calculated by:

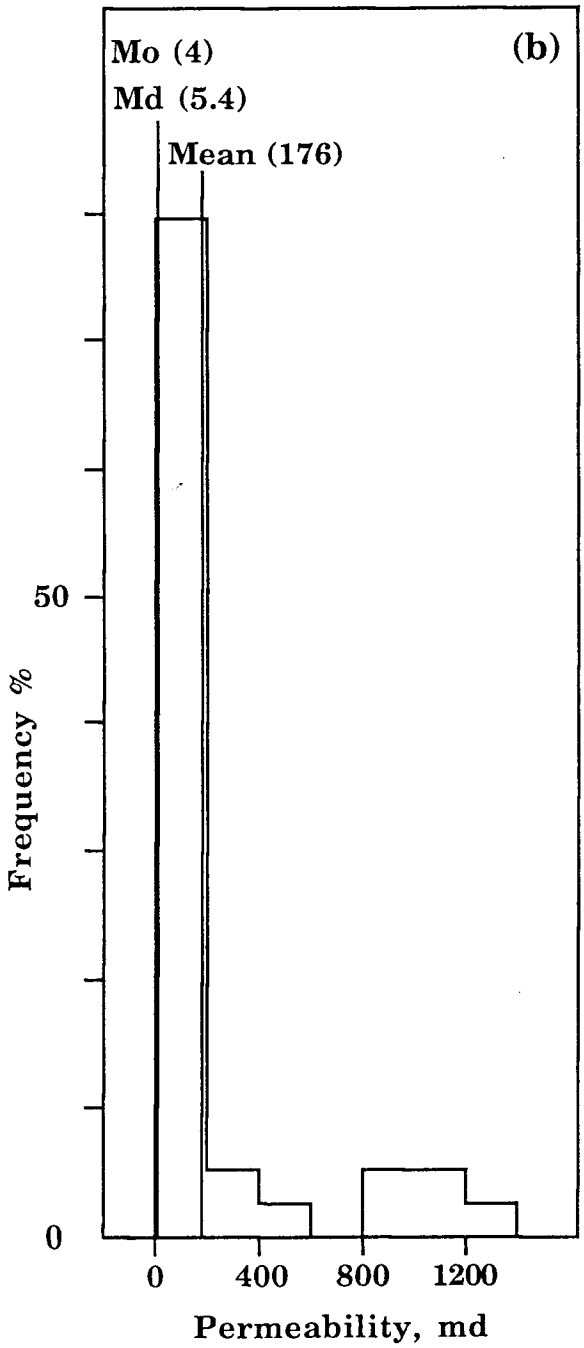
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

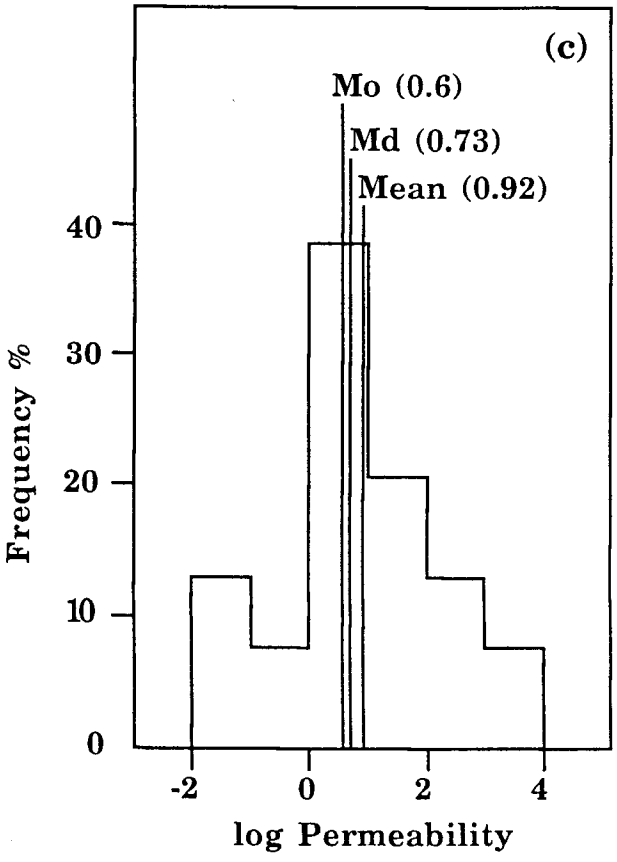
where n is the number of observations. The sample mean is an estimate of the population mean, μ. (Remember that sample estimates are in Latin script, population parameters are written in Greek.) In mechanical terms, the mean is the center of gravity of the distribution. Statistically, the mean is the expected value of the distribution and often written as E(X). The mean is the most sensitive measure of centrality. However, its real value over the mode and the median is that it is a fundamental measure used in a variety of parametric statistical inference tests.

The three central measures are shown for porosity and permeability distributions in the Picaroon sandstones (see DESCRIPTORS OF LOCATION). Notice in (a) how the mean, mode, and median almost coincide for porosity values, which is the expectation for symmetric distributions. The strongly asymmetric (positive skewed) distribution of permeabilities (b) results in a mean much higher than the mode or median. If the permeabilities are plotted and averaged in logarithmically-transformed units (c), the distribution is more symmetrical and the central measures are similar. The average of the logarithmic values corresponds to the geometric average of the raw values: $\overline{X}g = \sqrt[n]{\Pi X_i}$.

DESCRIPTORS OF LOCATION:
MEAN, MEDIAN, AND MODE OF PICAROON SANDSTONE
(a) POROSITY, (b) PERMEABILITY, AND
(c) LOGARITHMICALLY SCALED PERMEABILITY







DESCRIPTORS OF DISPERSION (measures of variability)

We have already seen the use of the interquartile range (from the 25% quartile to the 75% quartile) as a measure of the spread of a distribution. This is the appropriate statistic of dispersion to use when the median has been selected to represent the central location.

When the mean value is used as for the central measure, variability of the distribution about the mean is computed as the variance. For a complete population of observations, the variance is given by:

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{n}$$

In the more common case of a sample, the sample variance is an estimate of the population parameter and is computed by:

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{(n-1)}$$

Notice how the divisor is (n-1) because a degree of freedom was lost when the sample estimate of the mean was used.

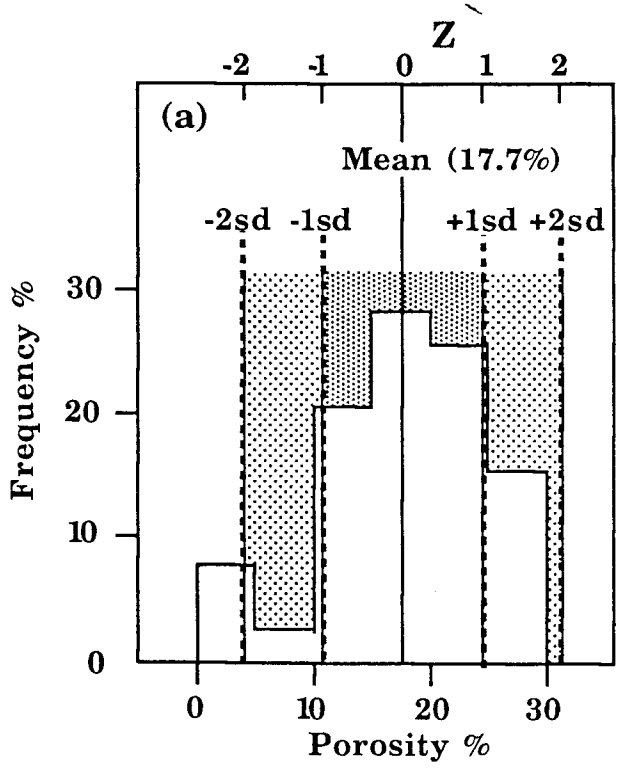
Variance is expressed in squared units, so a more tractable measure of spread is given by the standard deviation, s, which is the square root of the variance. The spread of a distribution can then be summarized as multiples of standard deviation distances about the mean value. This leads to a useful measure of the relationship between an observation, X_i, and the mean by:

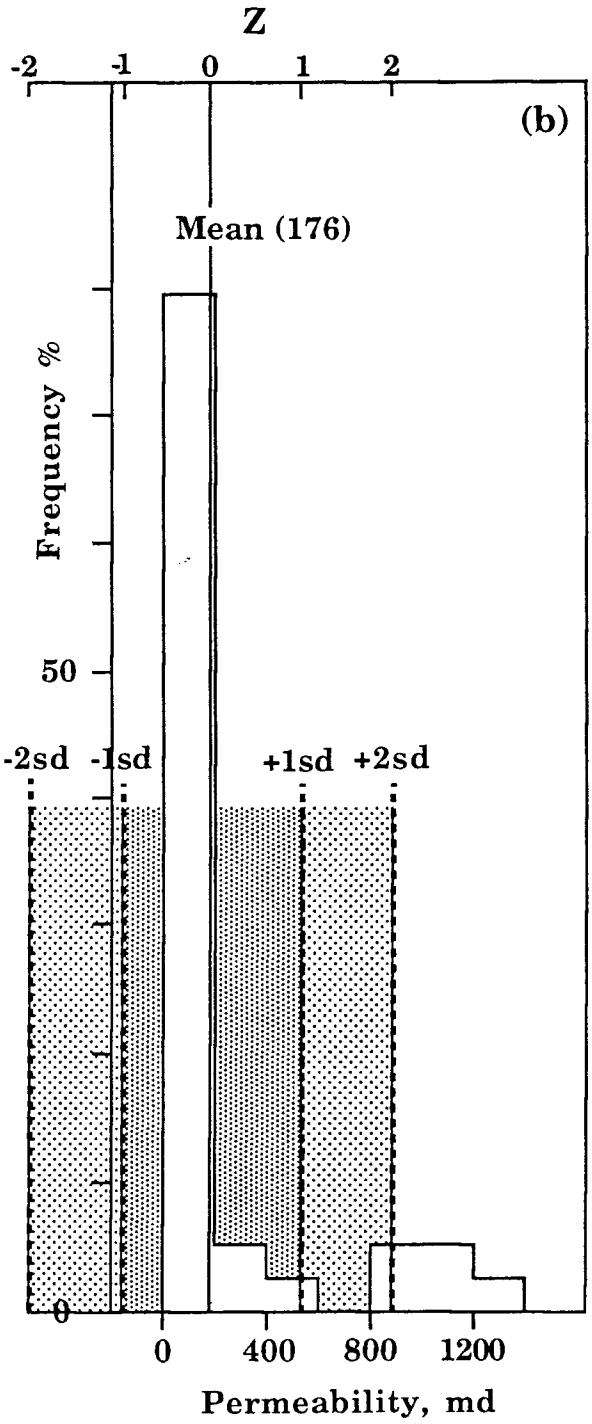
$$Z_i = \frac{X_i - \overline{X}}{s_X}$$

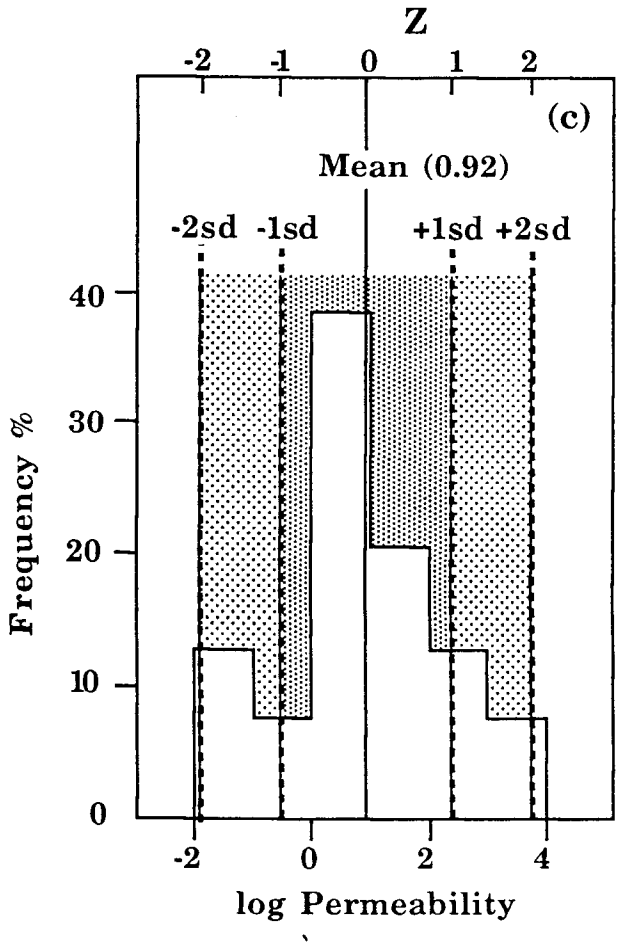
where Z_i is a Z-score or standardized score, and gives the distance of the observation to the mean in standard deviation units. The transformation allows us to recognize immediately both typical and extreme values. If the data follow a normal distribution, then our expectation is that 68% of the observations should fall within a range of plus or minus one standard deviation about the mean; 95% should occur within two standard deviation units from the mean.

Plots of the standard deviation ranges of porosity, permeability, and logarithmic permeability of the Picaroon sandstones illustrate these concepts (see STANDARD DEVIATIONS ...). The porosities and logarithmically-scaled permeabilities are fairly symmetrically distributed and their overall distribution character is captured by the mean and standard deviation ranges. The raw permeabilities are strongly skewed so that the standard deviation gives a poor representation of distribution dispersion. The selection of a useful scale transformation (most commonly logarithmic) will often remedy the problem, so that sample statistics are reasonable representations of the location and dispersion of distributions.

STANDARD DEVIATION RANGES ABOUT MEAN OF PICAROON SANDSTONE
(a) POROSITY, (b) PERMEABILITY, AND
(c) LOGARITHMICALLY SCALED PERMEABILITY







HIGHER ORDER MOMENTS (skewness and kurtosis)

The mean and variance are measures of location and dispersion respectively, and are the first two moments of a distribution. Recall that the estimate of the second moment about the mean, the variance, was calculated from:

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{(n-1)}$$

The third moment about the mean is the skewness, given by:

$$m_3 = \frac{\Sigma(X_i - \overline{X})^3}{(n-1)}$$

The value is in cubed units, so a dimensionless measure can be computed from:

$$Sk = \frac{n}{(n-1)(n-2)} \sum \left(\frac{X_i - \overline{X}}{s} \right)^3$$

If the skewness is zero then the distribution is symmetrical. Otherwise, a positive value of skewness shows a tendency for a right skew of an extended tail to positive values. A negative or left skew indicates a tail stretched towards lower values.

The fourth moment is kurtosis, given by:

$$m_4 = \frac{\Sigma(X_i - \overline{X})^4}{(n-1)}$$

with its dimensionless equivalent calculated by:

$$Kt = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{X_i - \overline{X}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

The kurtosis measures the relative "peakedness" of the distribution. More peaked distributions are "leptokurtic"; flatter distributions are "platykurtic". More realistically, the useful information is generally reflected in the corresponding contraction or stretching of the tails, rather than the peakedness of the central mode. The normal distribution is used as the reference distribution to make these assessments. A normally distributed distribution should generate a value of zero in the dimensionless equation shown above.

The dimensionless measures of skewness and kurtosis of the Picaroon sandstone porosity and permeability data are:

POROSITY:	Sk= -0.6	Kt= 0.1
PERMEABILITY:	Sk= 2.0	Kt= 2.5

The porosity distribution skewness and kurtosis are close to zero and so therefore similar to expectations of a normal distribution. The permeability distribution is leptokurtic and with a strong positive skew.

THE NORMAL DISTRIBUTION

The normal distribution is the bell curve that is familiar even to those who know little or nothing about statistics. What is it? Where did it come from? Why is it important?
The normal distribution is the limiting continuous form of the discrete binomial distribution. The binomial distribution applies to events with two possible outcomes and describes the number of "successes" (or "failures") that occur within a given number of trials. If the probability of success is p, and the probability of failure is q $(=1-p)$, then the proportion of r successes in n trials is given by:

$$P(r) = \frac{n!}{(n-r)!r!} q^{n-r} p^r$$

The sample mean and variance of the number of successes are:

$$\overline{X} = np \quad \text{and} \quad s^2 = npq$$

Obviously, the possible values of r must range between zero and n, so that the results can be shown as a bar graph for each incremental value of possible r (see BINOMIAL ...).

As n becomes larger, the distribution grows progressively smoother until in the continuous limit (an infinite number of trials), it becomes the normal distribution (see BINOMIAL..). De Moivre showed this in the 17th century as a theoretical result but saw no practical use for it. The formula for the standard normal distribution is:

$$P(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)}$$

Notice that the distribution is defined completely by the mean (μ) and the standard deviation (σ). The "standard normal distribution" has a mean of zero and a standardized standard deviation of one. Remember that this can be produced from any measurement scale by a Z-score transformation:

$$Z = \frac{(X_i - \overline{X})}{s} \text{ for a sample, or } Z = \frac{(X_i - \mu)}{\sigma} \text{ for a population.}$$

Z is the mean and variance term in the normal distribution formula exponent. When data are normally distributed, approximately 68% will lie within a range of plus or minus one standard deviation from the mean; 95% for two standard deviations (see NORMAL DISTRIBUTION). Most statistics text contain a normal distribution function table which relates Z-scores to the corresponding proportional area under the curve (see NORMAL DISTRIBUTION FUNCTION). Later, we shall see how this information is used in statistical inference tests.

It was Gauss who first recognized that the normal distribution could be used to model the distribution of random errors about a hypothetical true value. Consequently the distribution is often known as the Gaussian curve or normal

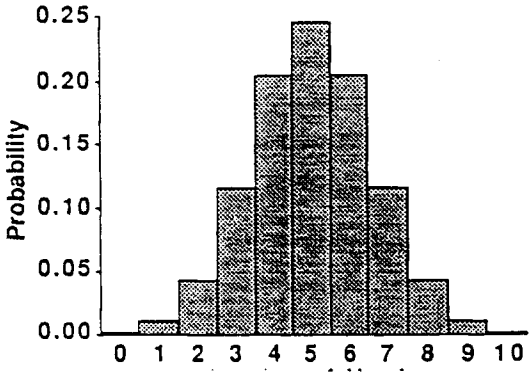
law of error. Conceptually, measurement errors are considered to result from the arithmetic summation of many small positive and negative random incremental displacements from a mean. This is a binomial model which generates a normal distribution in the limit on a continuous measurement scale.

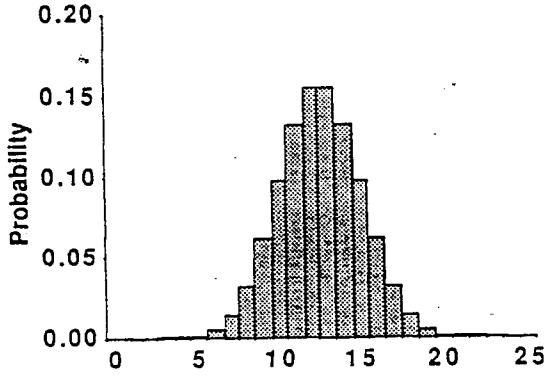
The normal distribution was proposed as a theoretical model for the analysis of statistical error (and still is used for that purpose). However, natural variabilities (height, weight, porosity) often took the form of a more-or-less symmetrical clump of values that faded into tails of less frequent and extreme values and could be fitted adequately by a normal distribution. A theoretical justification of this match can become more difficult because it implies that there is a "natural" value for many measurements and that observations that deviate from this ideal are "errors". However, irrespective of whether we "believe" in this interpretation, a close match of a normal distribution to the observed variability is a very useful result. This is because the normal distribution is completely defined by its mean and variance. Therefore, we will have captured almost the entire variability of the observations with just two numbers. Of course, our calculations of these statistics will be only *estimates* of their true population *parameters*. However, a reasonable match will allow us to use the power of parametric inferential statistics to make conclusions concerning similarities and differences, correlations, and predictions. These are the methods of classical statistics which we will review later in this manual.

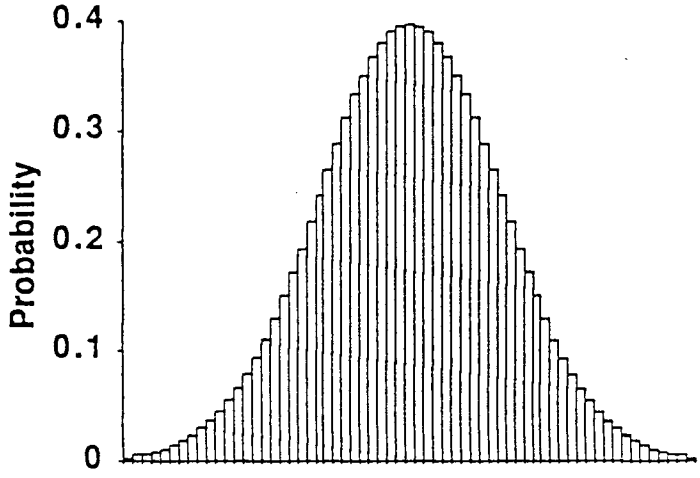
Earlier, we estimated the means and standard deviations of porosities, permeabilities, and logarithmically-scaled permeabilities in the Picaroon sandstones. These estimates can be used to generate hypothetical normal distributions for visual comparisons with actual variability (see FITTED NORMAL DISTRIBUTION ..). Notice the good fit to the porosity data (a) and the poor fit to the permeabilities (b). (A review of the skewness and kurtosis estimates of these data sets would have alerted us to this result before we plotted any graphics.) However, there is a marked improvement in fit when the normal distribution is modeled on logarithmically-scaled porosities (c). This implies that the logarithms of the observations may be normally distributed. If this is indicative of some natural process then the driving model is a multiplicative one, rather than the additive model used by Gauss for random errors. The lognormal distribution has been widely used to represent variation caused by processes such as breakage or agglomeration. So, it is both a reasonable theoretical model and often a good empirical match to petroleum geology variables such as permeabilities and field sizes.

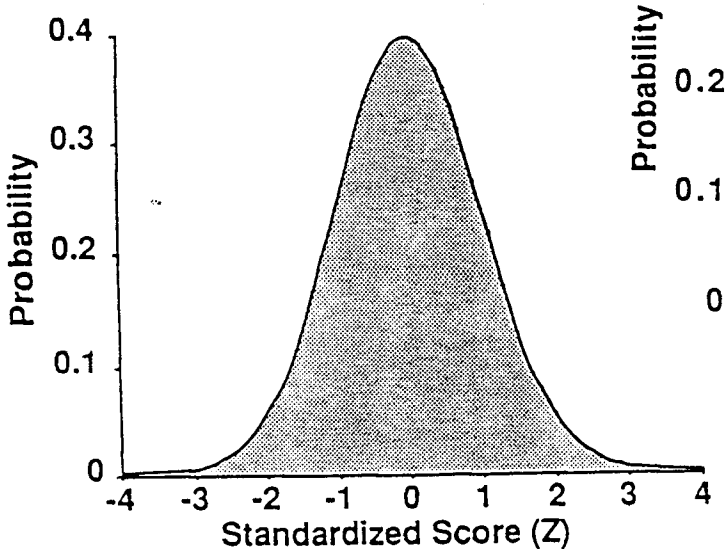
The degree of fit to a normal distribution can also be assessed by plotting observations on a normal probability plot (or P-plot). The convention is the same as a Q-plot, but the cumulative probabilities are scaled to Z-score values (see NORMAL PROBABILITY PLOT). If normal, the points should plot on a straight line. P-plots are also available for lognormal and other hypothetical distributions.

BINOMIAL TO NORMAL DISTRIBUTION
Transition from a discrete to a continuous distribution **p=0.5**

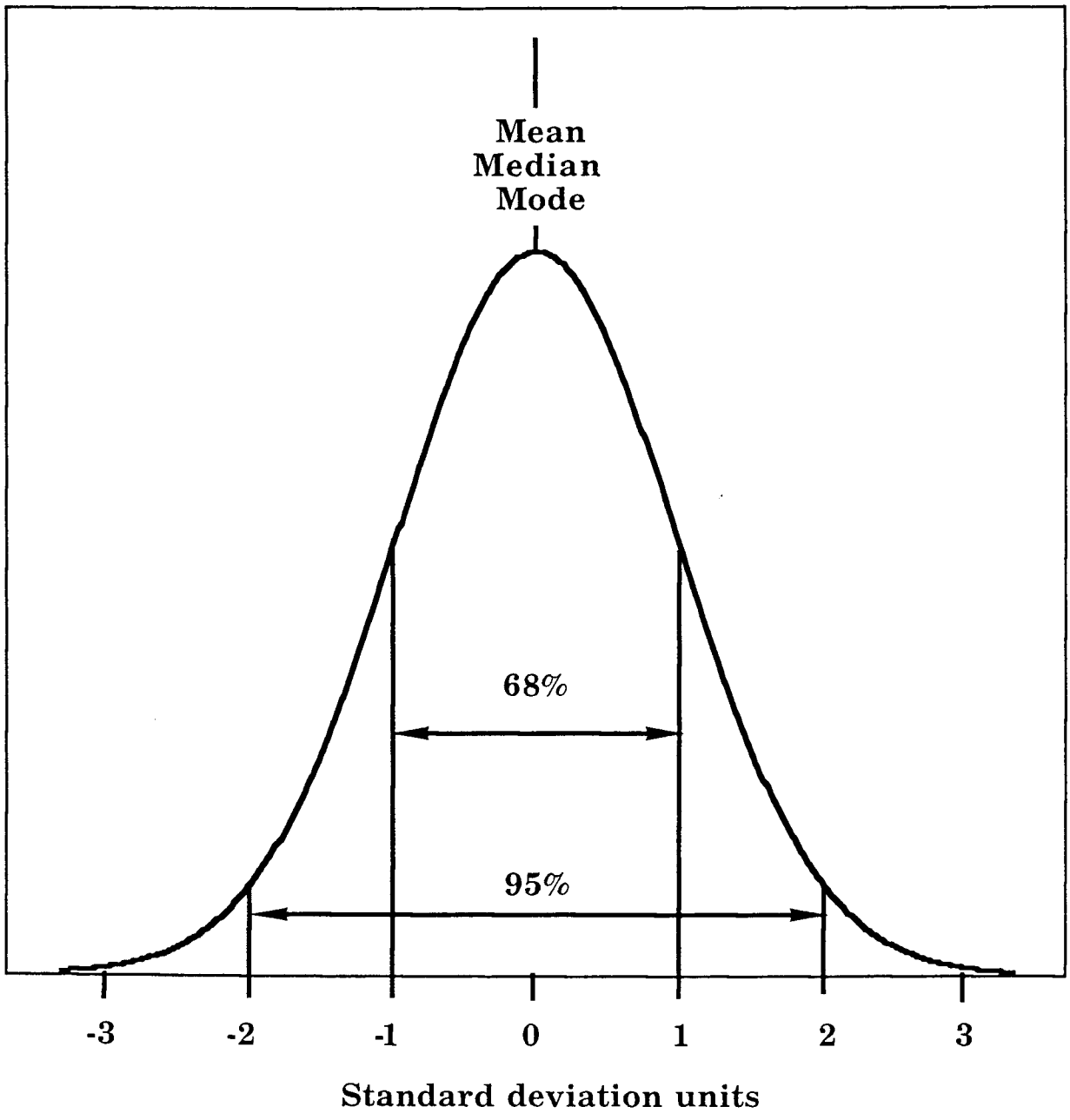








THE NORMAL DISTRIBUTION



NORMAL DISTRIBUTION FUNCTION

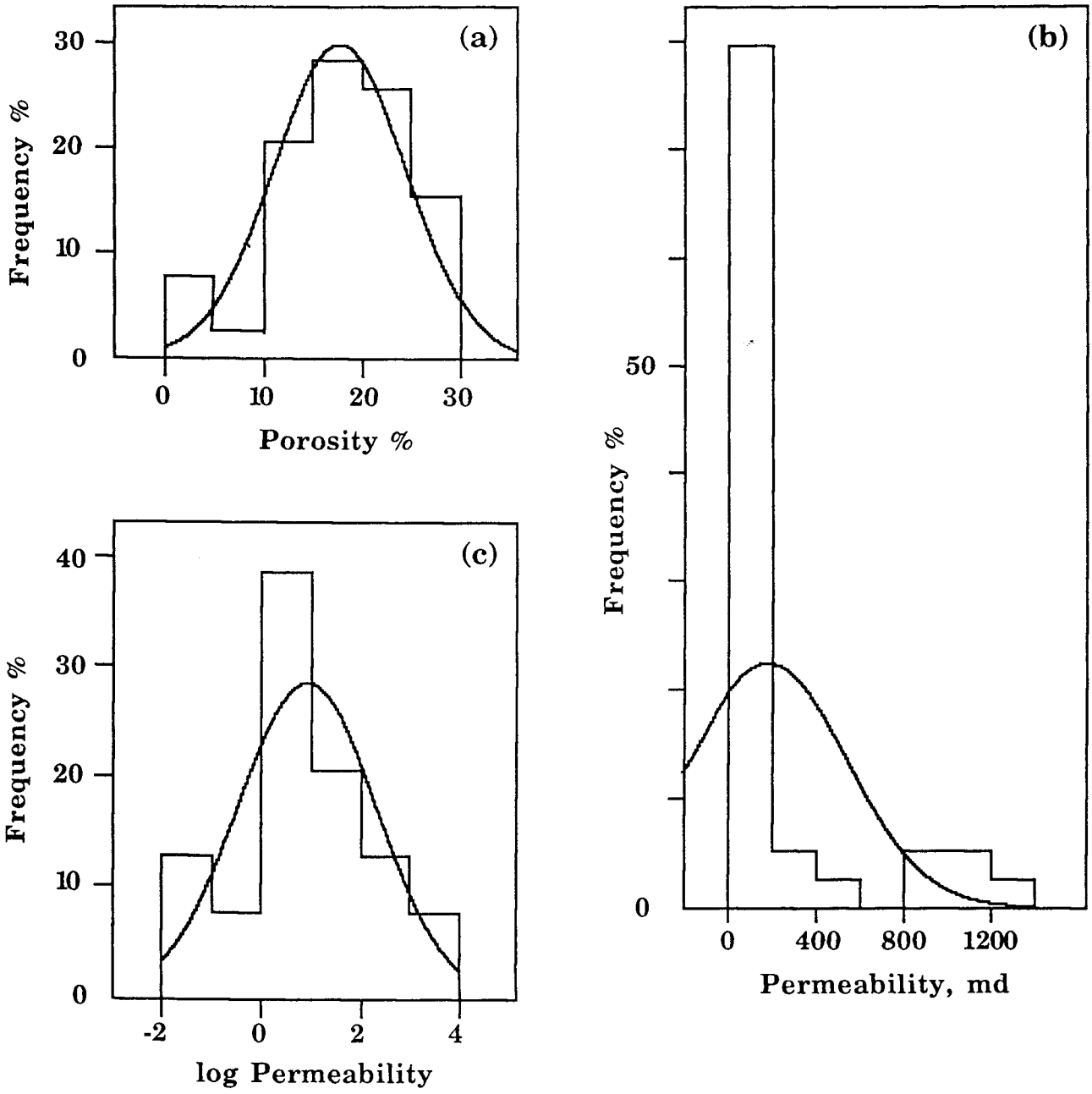
Area under the normal curve.

The area given is

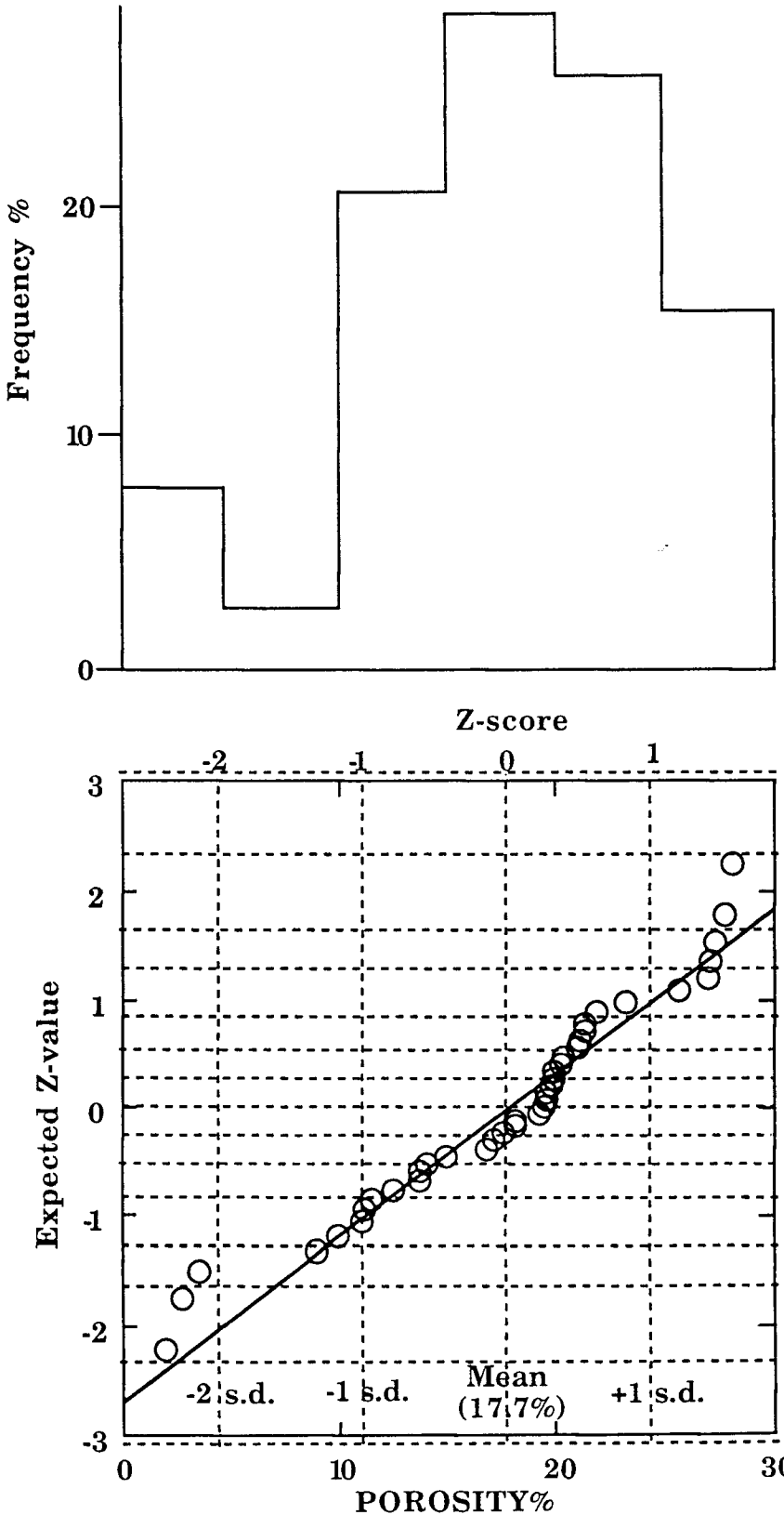
$$\int_{-\infty}^{z} \frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}z^2} dz.$$

	0·00	0·01	0·02	0·03	0·04	0·05	0·06	0·07	0·08	0·09
0·0	0·5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0·1	0·5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0·2	0·5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0·3	0·6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0·4	0·6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0·5	0·6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0·6	0·7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0·7	0·7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0·8	0·7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0·9	0·8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1·0	0·8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1·1	0·8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1·2	0·8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1·3	0·9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1·4	0·9192	9207	9222	9236	9251	9265	9279	9292	9306	9319
1·5	0·9332	9345	9357	9370	9382	9394	9406	9418	9429	9441
1·6	0·9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1·7	0·9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1·8	0·9641	9649	9656	9664	9671	9678	9686	9693	9699	9706
1·9	0·9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
2·0	0·9772	9778	9783	9788	9793	9798	9803	9808	9812	9817
2·1	0·9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2·2	0·9861	9864	9868	9871	9875	9878	9881	9884	9887	9890
2·3	0·9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2·4	0·9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2·5	0·99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2·6	0·99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2·7	0·99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2·8	0·99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2·9	0·99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3·0	0·99865	99869	99874	99878	99882	99886	99889	99893	99897	99900
3·1	0·99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3·2	0·99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3·3	0·99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3·4	0·99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3·5	0·99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3·6	0·99984	99985	99985	99986	99986	99987	99987	99988	99988	99989
3·7	0·99989	99990	99990	99990	99991	99991	99992	99992	99992	99992
3·8	0·99993	99993	99993	99994	99994	99994	99994	99995	99995	99995
3·9	0·99995	99995	99996	99996	99996	99996	99996	99996	99997	99997

NORMAL DISTRIBUTION CURVES FITTED TO PICAROON SANDSTONE HISTOGRAMS OF (a) POROSITY, (b) PERMEABILITY, AND (c) LOGARITHMICALLY SCALED PERMEABILITY



NORMAL PROBABILITY PLOT (or Normal P-Plot) OF PICAROON SANDSTONE POROSITY



THE CENTRAL LIMIT THEOREM

This most important theorem states that if random samples are drawn from a population, their means will tend to be normally distributed (more so for larger sample sizes) regardless of whether the measurement is itself normally distributed. The theorem is the basis for establishing confidence intervals around estimates and for important inference tests concerning potential similarities or differences between samples.

The sample means can be written as $\overline{X}_1, \overline{X}_2, \ldots\ldots$ and are scattered about a hypothetical population mean of μ. If the samples are small in size, then the sample means will be widely scattered about the population mean ; if very large, then the sample means will show little difference from the population parameter. Now, the standard deviation of individual observations about the population mean is the parameter, σ, which we are usually forced to estimate from a sample calculation of s. The standard deviation of sample means with n observations about the population mean is called the standard error of the mean and is given by:

$$s_e = \frac{\sigma}{\sqrt{n}}$$

Of course, we generally do not know the population standard deviation, so we substitute the sample estimate of s to give:

$$s_e = \frac{s}{\sqrt{n}}$$

As a practical example of the calculation of the standard error and how we use it, let us consider the Picaroon sandstone porosities. Now, the average porosity is 17.7% which is an estimate of the hypothetical Picaroon sandstone population average but is based only on 39 observations. The standard error of this estimate is then:

$$s_e = \frac{6.7}{\sqrt{39}} = 1.07$$

Just as standard deviation units define proportions of expected observations about a sample mean, standard error units will define proportions of expected sample means about the true population mean. Theoretically, 95% of sample means should be found within a distance of 1.96 standard errors from the true mean. Turning this around, we can say that we are 95% confident that the true mean lies within 1.96 standard errors of our estimate. If 95% confidence is a level with which we are comfortable we define a 95% confidence interval of:

$$\overline{X} \pm 1.96 s_e = 17.7 \pm 2.10\%$$

In other words, we are 95% confident that the true average porosity is somewhere between 15.6 and 19.8%.

If 95% confidence was considered too risky and a 99% confidence level was selected, then the critical number of standard error units would increase to 2.58 and the confidence interval of the population mean would expand to a range between 14.9% and 20.5%.

Of course, these estimates are based on the assumption that the 39 Picaroon sandstone porosity measurements collectively represent an unbiased, random sample of the hypothetical population of all Picaroon sandstones. The calculations could have some important economic consequences. Say, for example, the measurements were used to estimate the average porosity of the entire Picaroon field. We will further suppose that the pay zone has been defined as the sandstones of "Unit 1" and that the statistically-intelligent engineering measurement have requested a 99% confidence interval to contain the true field average porosity.
Now, the mean value is 18.1% and the standard deviation of "Unit 1" porosities is 3.4%, based on 20 samples. Then the standard error, s_e, is 0.76%. This is less than the standard error for the larger sample but the decrease is caused by the marked drop in standard deviation within "Unit 1". Then the confidence interval at 99% is:

$$18.1 \pm 2.58 s_e = 18.1 \pm 1.96\%]$$

Let us next suppose that management has declared this result to be unacceptable: they need to have a 99% confidence interval of plus or minus one porosity unit. While we cannot improve on our estimate based on this sample, we can make an estimate of how many core sample measurements of Picaroon sandstone "Unit 1" we would need to satisfy this demand. The required standard error is:

$$s_e = \frac{1}{2.58} = 0.39\%$$

Using our standard deviation estimate of 3.4%, the estimated number of core samples needed is:

$$n = \left(\frac{s}{s_e}\right)^2 = \left(\frac{3.4}{0.39}\right)^2 = 76$$

Notice how the philosophy of statistical inference can be used both to analyze existing data and as a methodology to **plan** work so that results can be estimated at preset levels of confidence. This is the hallmark of classical statistics which concerns itself with **both** the design and analysis of experiments.

STUDENT'S t-DISTRIBUTION

W.S. Gossett reported pioneer statistical work under the penname of "Student", because at that time his employers, the Guiness Brewery of Dublin, did not allow its staff to publish research openly. Student worked with small observation samples from yeast counts in brewing. As already discussed, the Central Limit Theorem states that sample means will tend to be normally distributed about the population mean with a standard deviation equal to the standard error. If this normal distribution is plotted out in standardized form (i.e. with a mean of zero and a standard deviation of one), then this is a distribution of Z-scores, where:

$$Z = \frac{(\bar{X} - \mu)}{s_e}$$

The standard error is given by:

$$s_e = \frac{\sigma}{\sqrt{n}}$$

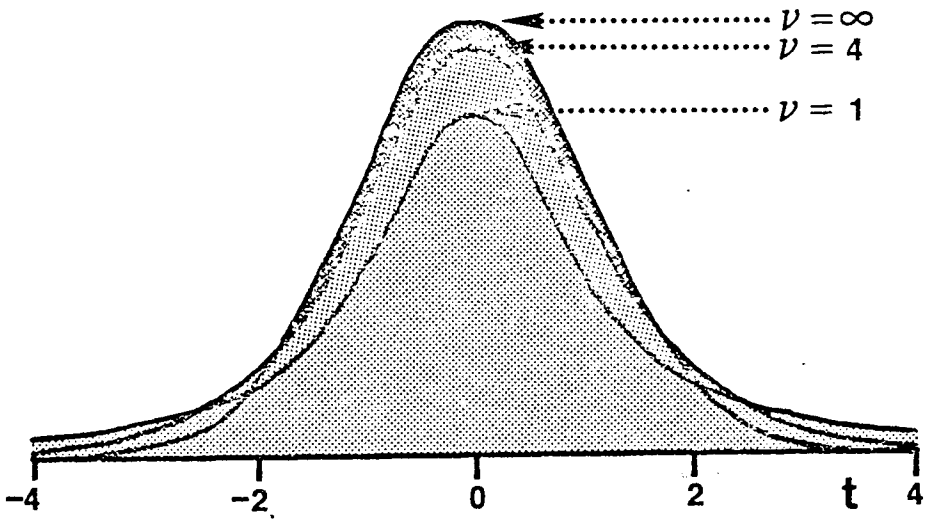
As we noted before, we generally do not know the population standard deviation, σ, so we substitute the sample estimate of s to give:

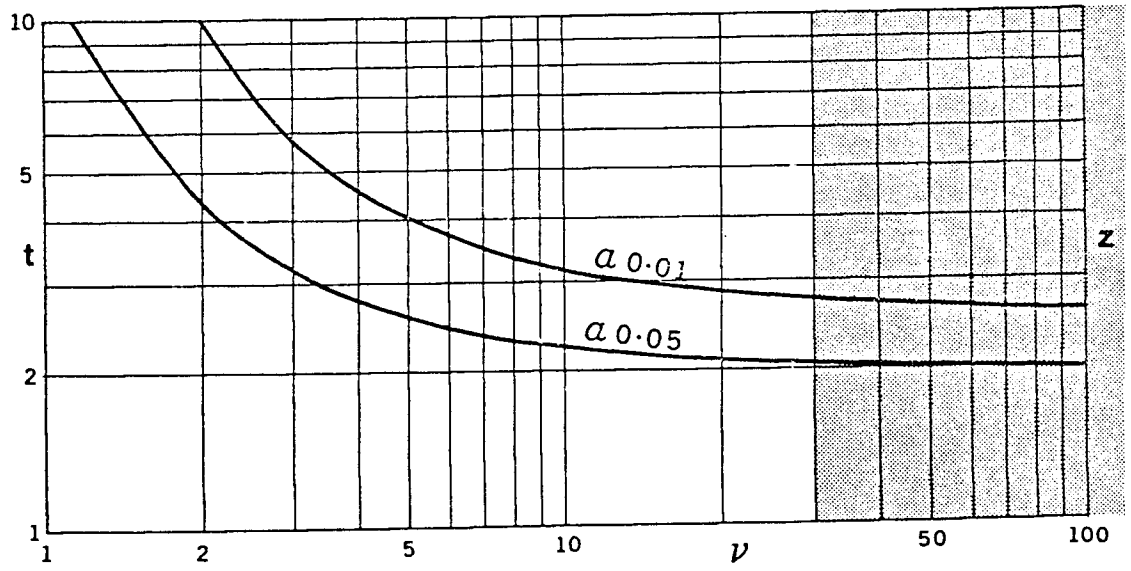
$$s_e = \frac{s}{\sqrt{n}}$$

As the number of observations in a sample becomes larger, s gets closer to σ, so that the difference becomes negligible. However, at small sample sizes, s is not only a poor estimate, but it is biased: it is an underestimate of σ. The reason for this is that the sample standard deviation is related to the *sample* mean rather than the *population* mean. The squared differences of observations to any value other than their common mean will always be larger. Therefore, when the Z-scores for sample means are calculated for small samples, using the sample standard deviations, the resulting distribution is broader than the normal, and is known as the t-distribution.

The shape of the t-distribution changes with the sample size, n. As n grows larger, the distribution contracts until it converges on the normal distribution (see t-DISTRIBUTION). In practice, it is difficult to tell the t-distribution from the normal at sample sizes greater than about 30. This gives a useful rule of thumb: "small samples" are often considered to be those with less than about 30 observations; "large samples" are those with more than 30. The t-distribution is tabulated in most statistics texts (see t-TABLE) for use in the Student t-test that we will examine later in this manual. Notice that with larger samples, the values of t approach those of Z from the standard normal distribution.

THE t-DISTRIBUTION





THE t-TABLE

ν	Probability 0·05	0·02	0·01	0·001
1	12·706	31·821	63·657	636·619
2	4·303	6·965	9·925	31·598
3	3·182	4·541	5·841	12·941
4	2·776	3·747	4·604	8·610
5	2·571	3·365	4·032	6·859
6	2·447	3·143	3·707	5·959
7	2·365	2·998	3·499	5·405
8	2·306	2·896	3·355	5·041
9	2·262	2·821	3·250	4·781
10	2·228	2·764	3·169	4·587
11	2·201	2·718	3·106	4·437
12	2·179	2·681	3·055	4·318
13	2·160	2·650	3·012	4·221
14	2·145	2·624	2·977	4·140
15	2·131	2·602	2·947	4·073
16	2·120	2·583	2·921	4·015
17	2·110	2·567	2·898	3·965
18	2·101	2·552	2·878	3·922
19	2·093	2·539	2·861	3·883
20	2·086	2·528	2·845	3·850
21	2·080	2·518	2·831	3·819
22	2·074	2·508	2·819	3·792
23	2·069	2·500	2·807	3·767
24	2·064	2·492	2·797	3·745
25	2·060	2·485	2·787	3·725
26	2·056	2·479	2·779	3·707
27	2·052	2·473	2·771	3·690
28	2·048	2·467	2·763	3·674
29	2·045	2·462	2·756	3·659
30	2·042	2·457	2·750	3·646
40	2·021	2·423	2·704	3·551
60	2·000	2·390	2·660	3·460
120	1·980	2·358	2·617	3·373
∞	1·960	2·326	2·576	3·291

STATISTICAL HYPOTHESIS TESTS

The computation of the estimates of parameters such as means and standard deviations results in numbers that constitute "descriptive statistics". Conclusions may be drawn concerning similarities, differences, trends, and other patterns by visual inspection. However, interpretations will vary between individuals and any conclusions will ultimately be subjective. "Inferential statistics" are a battery of techniques that enable workers to arrive at decisions in a consistent and rational manner. Unless the problem is trivial, any answer will have a certain degree of uncertainty associated with it. Therefore the conclusions are phrased in terms of probability. It is for the user to select the level of risk associated with the decision that he or she considers to be acceptable.

Statistical inference is a process of inductive logic: generalizations concerning a large population are made from the statistics of a limited sample. Statistical procedures follow the basic scientific method: a hypothesis is proposed and put to the test. However, statistical inference philosophy is very similar to that of the courtroom: motivated hypotheses are accepted providing null hypotheses are rejected at a convincing level of probability. Therefore, nothing is "proved", just as no-one is ever found "innocent" in a court of law, but may be found "not guilty". Conviction fails because insufficient evidence is presented to find a verdict of "guilty" beyond a reasonable doubt.

Statistical hypotheses

The null hypothesis of most statistical tests proposes that there is no difference between the population parameter estimated from the sample statistics of one or more samples. In other words, that the samples are from a common population. So, for example, if we were interested in whether there was a significant difference in the estimates of the means of two samples, we would write the null hypothesis formally as:

$$H_0: \mu_1 = \mu_2$$

The alternative (and motivated) hypothesis is:

$$H_1: \mu_1 \neq \mu_2$$

A test criterion is set for the rejection of the null hypothesis. The criterion is based on probability and is chosen to reflect the cost of being wrong. The appropriate test statistic is computed and compared with the criterion value. The null hypothesis is either accepted or rejected as a result of the comparison. If the null hypothesis is rejected the alternative hypothesis is accepted.

Significance levels

A significance level is the acceptable risk of making a Type I error: the probability that the null hypothesis is really true even when it was rejected by the statistical test. The significance level is denoted by α and should be selected *before* the test. This convention cuts down on gerrymandering conclusions by relaxing probability levels in favor of a desired result when it fails to make the cut. Most software packages deliver the statistic and the matching significance level that the statistic will pass in order to reject the null hypothesis. So, there is an opportunity for post-analysis rationalization to prejudice an α level in favor of a desired result. In most geological problems, a value of 0.05 is customarily used, in common with many scientific applications. Probability assignments can be less intuitive when there is a monetary or other tangible risk involved. Type II error is the probability of accepting a hypothesis when it is false. This probability is generally unknown, so we attempt to minimize it by setting the null hypothesis as the one we are attempting to reject. By using a conservatively low level of α, the chance of a Type II error is also reduced.

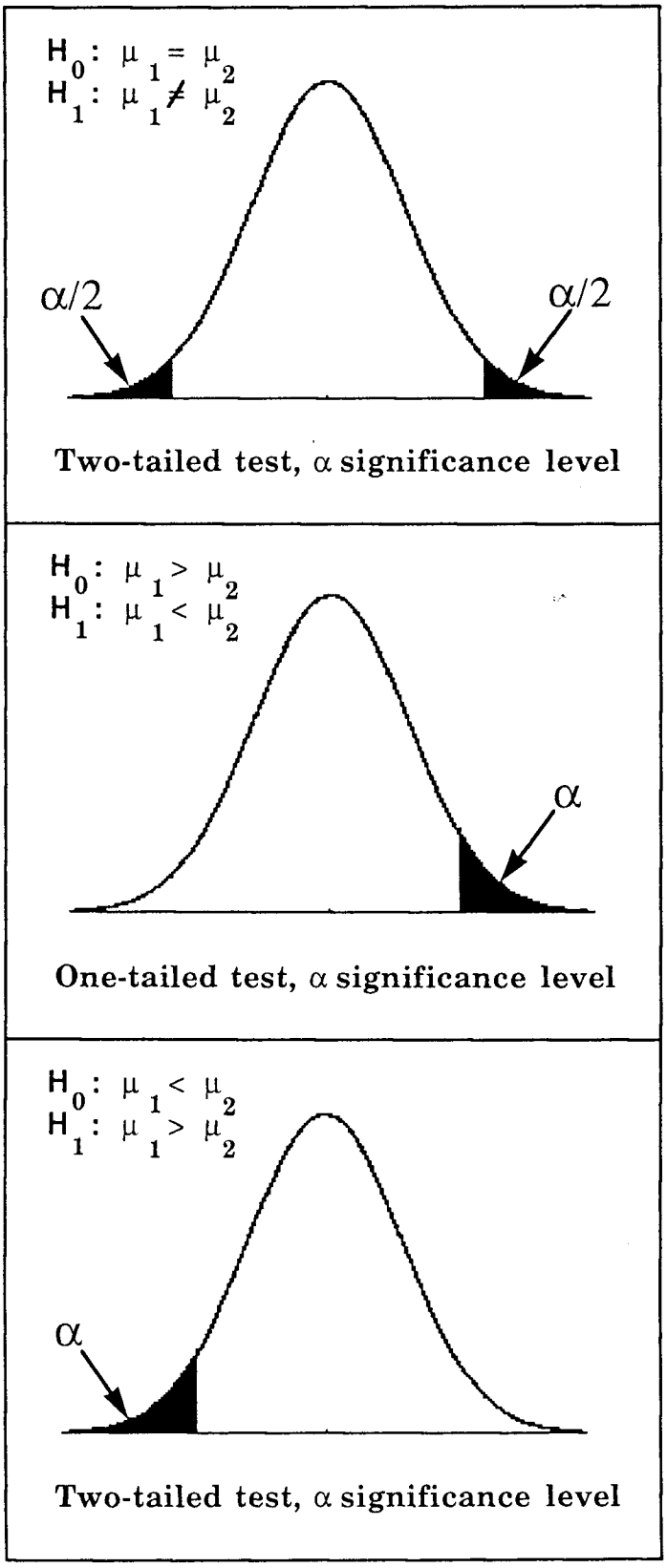
Directional and non-directional hypotheses

The null hypothesis of equivalence of means implies a *non-directional* motivated hypothesis that the second mean could be either more or less than the first mean. If the purpose of the test was to evaluate whether the second mean was significantly *greater* than the first, then the motivated hypothesis would be *directional*. Then the null hypothes would be modified to the proposition that the second mean is either equal or less than the first. A *two-tailed test* is used for non-directional hypotheses; a *one-tailed test* for directional hypotheses (see DIRECTIONAL AND NON-DIRECTIONAL HYPOTHESES). If a 5% significance level is selected, then the sample statistic must lie in the 5% extreme of the distribution. However, in the one-tailed case, the 5% is in one tail; in the two tailed case it is the sum of two 2.5% components in each tail. So, whether the test is one- or two-tailed will result in a different choice of critical statistic, even for the same significance level.

Degrees of freedom

Each time we estimate a parameter, we lose a degree of freedom. In effect, the number of degrees of freedom is the number of independent items of information in a sample. Usually, the number of degrees of freedom is equal to the number of observations in the sample minus the number of parameters that have been estimated. The degrees of freedom are often either abbreviated as 'df' or symbolized as ν. The degrees of freedom must be enumerated for any hypothesis test, because the critical test value will be set by the number of degrees of freedom and the selected significance level, α.

DIRECTIONAL AND NON-DIRECTIONAL HYPOTHESES



THE t-TEST

The t-test is widely used to test the hypothesis of the equality of means:

$$H_0: \mu_1 = \mu_2$$

The alternative (and motivated) hypothesis is:

$$H_1: \mu_1 \neq \mu_2$$

In other words, is the difference between the estimate of the means of two samples sufficient to reject the hypothesis that they are two samples from the same population? If the difference is significant, then we can conclude that the two samples come from two different populations whose parameter means are different.

The t-statistic is computed from:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{s_e}$$

where $\overline{X}_1, \overline{X}_2$ are the sample means and s_e is the standard error.

To clarify these ideas, we will work through a t-test example, using porosities of the Picaroon sandstones. The distributions of porosities in "Unit 1" and "unit 2" look different (see COMPARISON OF POROSITY DISTRIBUTIONS...), but are the differences significant, when considering the small sample sizes involved? The statistics that we need for the two groups to make a t-test are:

UNIT 1:	$n_1 = 20$	$\overline{\Phi}_1 = 18.1\%$	$s_1 = 3.4\%$
UNIT 2:	$n_2 = 8$	$\overline{\Phi}_2 = 13.4$	$s_2 = 6.1\%$

The classical t-test can be applied when the variance of the two groups is essentially the same. The sample standard deviations show clearly that this is not the case. So, a *modified* t-test must be used. The best estimate of s_e, the standard error of the means is given by:

$$s_e = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad = 2.29$$

Now, the test value for t can be calculated from the formula above as:

$$t = \frac{18.1 - 13.4}{2.29} = 2.05$$

In other words, the two sample means are separated by a little over two standard error distances.

The approximate form of the t-test leads to an unusually complex estimation of the number of degrees of freedom by:

$$\nu = \frac{\left(s_1^2 / n_1 - s_2^2 / n_2\right)^2}{\left(s_1^2 / n_1\right)^2 / (n_1 - 1) + \left(s_2^2 / n_2\right)^2 / (n_2 - 1)}$$

from which $\nu \approx 8.8$ and so ν is estimated to be 9 because it must be an integer. If we choose a significance level of 10%, then the α-level is 0.10. This decision means that we have accepted the contingency that if we reject the null hypothesis, then there is a one in ten chance that we could be wrong. The critical value of t is found from the t-table as:

Critical t value @ α 0.10 and 9 df = 1.833

(This critical value is for a *two-tailed test* in which the ten percent is allocated between the two tails of 5% each.)

The computed statistic exceeds this critical value and so we reject the null hypothesis and accept the alternative hypothesis that they are samples taken from different populations. The porosities of Units 1 and 2 appear to be significantly different. Notice that if we had elected to go with a 5% significance level, we would not have rejected the null hypothesis. Would this have "proved" that they were samples from a common population? No, it would not. Instead, it would show that the evidence was not sufficiently overwhelming to rule out the null hypothesis. The "hung jury" would be caused by several factors including the direction from the bench of a more stringent probability of reasonable doubt (5%), the relatively small samples available, and the strong difference in sample variances which caused the estimated degrees of freedom to plummet. If the sample variances were similar, then the classical t-test could have been used and the degrees of freedom would have been given by:

$$\nu = n_1 + n_2 - 2 \quad = 26$$

In this case, the standard error could have been estimated from pooling the samples in a common estimate of population variance as:

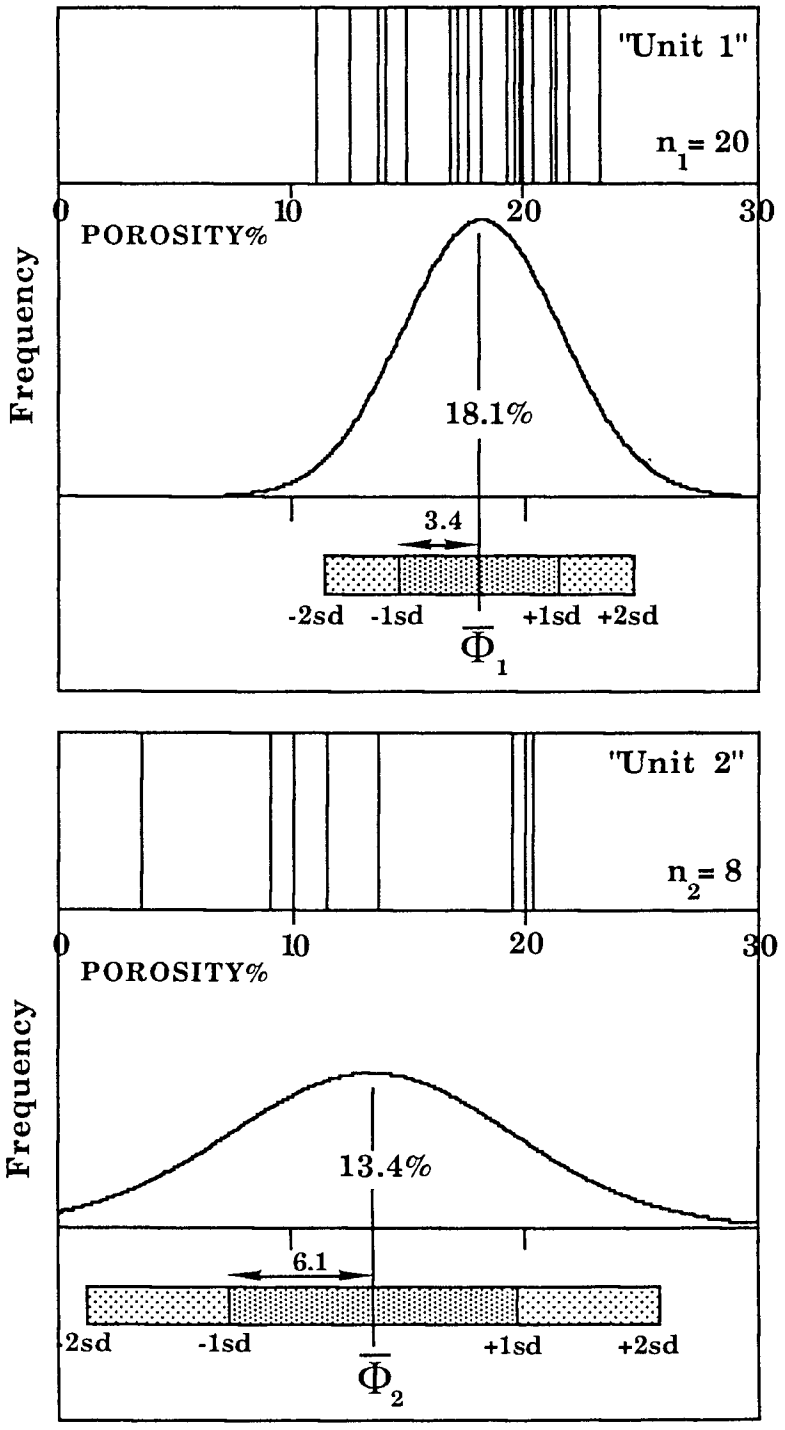
$$s_p^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

and computing:

$$s_e = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

How would we test to see if the sample variances were significantly different? This can be done with an F-test.

COMPARISON OF POROSITY DISTRIBUTIONS BETWEEN PICAROON SANDSTONE "UNIT 1" AND "UNIT 2"



THE F-TEST

The F-distribution describes the expected values of the ratio of the variances of two samples that have been taken from the same normal distribution:

$$F = \frac{s_1^2}{s_2^2}$$

The larger sample variance is placed as the numerator, the smaller as the denominator in the ratio computation. The form of the distribution is controlled by the number of observations in each sample, n_1 and n_2. The distribution is the basis for the F-test, which is widely used in the analysis of variance (ANOVA) and in regression.

The null hypothesis of an F-test is that variances calculated from two samples represent sample estimates of the same population variance:

$$H_0: \sigma_1^2 = \sigma_2^2$$

If this is true, then the ratio should be close to one, but there will be some variability because of sampling fluctuation. The question then becomes: "Is the ratio so large that I find it difficult to believe that the samples are from the same population?"

For a simple example, we can test the porosity variances in Units 1 and 2 of the Picaroon sandstones to see if they are significantly different. The necessary information is:

UNIT 1:	$n_2 = 20$	$s_2^2 = 11.4$
UNIT 2:	$n_1 = 8$	$s_1^2 = 37.3$

Remember that the sample with the larger variance is the numerator (and designated as Sample #1.) Then:

$$F = \frac{37.3}{11.4} = 3.27$$

The number of degrees of freedom to be used with this test are:

$$\nu_1 = n_1 - 1 = 7 \qquad \text{and} \qquad \nu_2 = n_2 - 1 = 19$$

Then the degrees of freedom are used to locate the test value in an F-distribution table (see F-TABLE):

Critical F-test value @ a 0.05 and 7 and 19 df = 2.54

The calculated value exceeds the test value and so the null hypothesis is rejected.

Later in the manual we will see that the F-test is a powerful tool to help determine significant components of regression models to be used in prediction.

F-TABLE

Five per cent points of the *F* distribution.

ν_1 must always correspond with the greater mean square.

The probability is 0·05 that a random variate should exceed *F*.

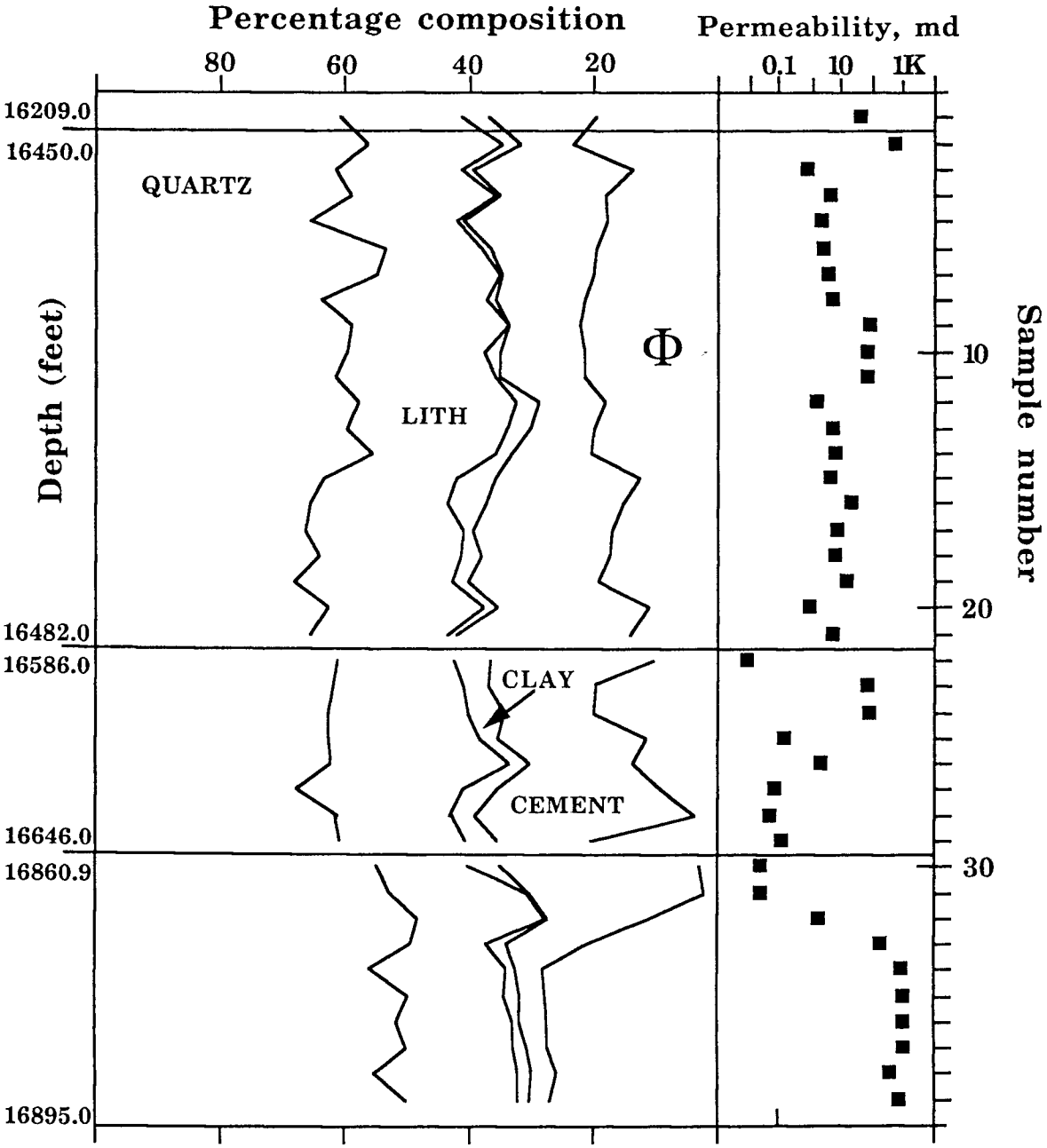
ν_1=	1	2	3	4	5	6	8	12	24	∞
ν_2=1	161·4	199·5	215·7	224·6	230·2	234·0	238·9	243·9	249·0	254·3
2	18·51	19·00	19·16	19·25	19·30	19·33	19·37	19·41	19·45	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·84	8·74	8·64	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6·04	5·91	5·77	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·82	4·68	4·53	4·36
6	5·99	5·14	4·76	4·53	4·39	4·28	4·15	4·00	3·84	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·73	3·57	3·41	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·44	3·28	3·12	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·23	3·07	2·90	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·07	2·91	2·74	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	2·95	2·79	2·61	2·40
12	4·75	3·88	3·49	3·26	3·11	3·00	2·85	2·69	2·50	2·30
13	4·67	3·80	3·41	3·18	3·02	2·92	2·77	2·60	2·42	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·70	2·53	2·35	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·64	2·48	2·29	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·59	2·42	2·24	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·55	2·38	2·19	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·51	2·34	2·15	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·48	2·31	2·11	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·45	2·28	2·08	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·42	2·25	2·05	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·40	2·23	2·03	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·38	2·20	2·00	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·36	2·18	1·98	1·73
25	4·24	3·38	2·99	2·76	2·60	2·49	2·34	2·16	1·96	1·71
26	4·22	3·37	2·98	2·74	2·59	2·47	2·32	2·15	1·95	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·30	2·13	1·93	1·67
28	4·20	3·34	2·95	2·71	2·56	2·44	2·29	2·12	1·91	1·65
29	4·18	3·33	2·93	2·70	2·54	2·43	2·28	2·10	1·90	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·27	2·09	1·89	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·18	2·00	1·79	1·51
60	4·00	3·15	2·76	2·52	2·37	2·25	2·10	1·92	1·70	1·39
120	3·92	3·07	2·68	2·45	2·29	2·17	2·02	1·83	1·61	1·25
∞	3·84	2·99	2·60	2·37	2·21	2·09	1·94	1·75	1·52	1·00

BIVARIATE ANALYSIS (applied to Picaroon sandstones composition, porosity, and permeability)

Up to this point, we have considered the graphing, descriptive statistics, and inference concerning variables taken singly. Collectively these are all *univariate* methods. There are many instances when associations between variables are important because they can give insight into causative processes or because the associations can be used for the purpose of making predictions. Methods that examine the interrelationships of two variables are termed *bivariate*. Two principal concerns of bivariate analysis are whether there is *correlation* between two variables and whether the value of one can be predicted on the basis of knowledge of the other, using *regression* analysis and related methods. Statistical inference is important as a means to make judgments about whether perceived correlations are significant, whether a prediction relationship is really useful, and the magnitude of errors that will be associated with a prediction.

The Picaroon sandstone data set will be used to demonstrate these procedures. The composition is tabulated in terms of quartz, lithic fragments, clay, cement, and porosity. Collectively, these variables form a closed system that sums to 100%. The composition of the Picaroon sandstones is plotted in depth order together with permeability as a profile (see COMPOSITION-PERMEABILITY PROFILE...).

COMPOSITION-PERMEABILITY PROFILE OF THE PICAROON SANDSTONES



COVARIANCE AND CORRELATION

Variance is used as a measure of univariate dispersion ; covariance is the equivalent measure of joint variation of two variables around their common mean. The equation for covariance between two variables, x and y, is given by:

$$\mathrm{cov}(x,y) = \frac{1}{(n-1)} \sum (X_i - \overline{X})(Y_i - \overline{Y})$$

Notice that the covariance of a variable to itself would be the same as the variance.

Just as the univariate normal distribution is defined uniquely by the mean and variance, the bivariate normal distribution is specified completely by the two means, the two variances and the covariance. Proportional contours of the bivariate normal take the form of ellipses and can be drawn on a bivariate scatter plot. The contours will show a good match with the density of the data cloud when the two variables are normally distributed. The contours show a poor fit to (a) porosity-permeability covariation in the Picaroon sandstones (see BIVARIATE NORMAL DISTRIBUTION ...) but a marked improvement in (b) when the permeabilities are scaled logarithmically.

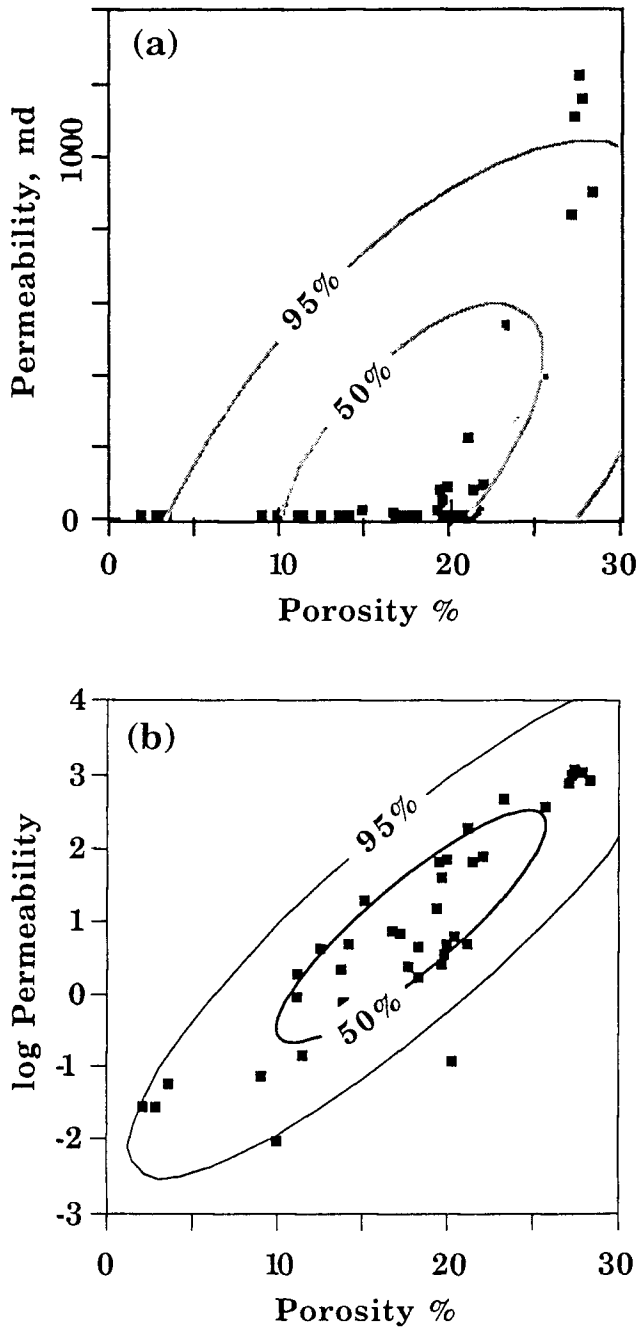
The Pearson product-moment correlation coefficient is a standardized measure of the linear relation between two variables. It is equivalent to the covariance of two variables after they have been standardized (by Z-score transformation) and so is independent of measurement units. When calculated from a sample, the correlation coefficient is an estimate, symbolized by r, of a population parameter coefficient, written as ρ. The equation for the Pearson correlation coefficient is:

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{s_x s_y}$$

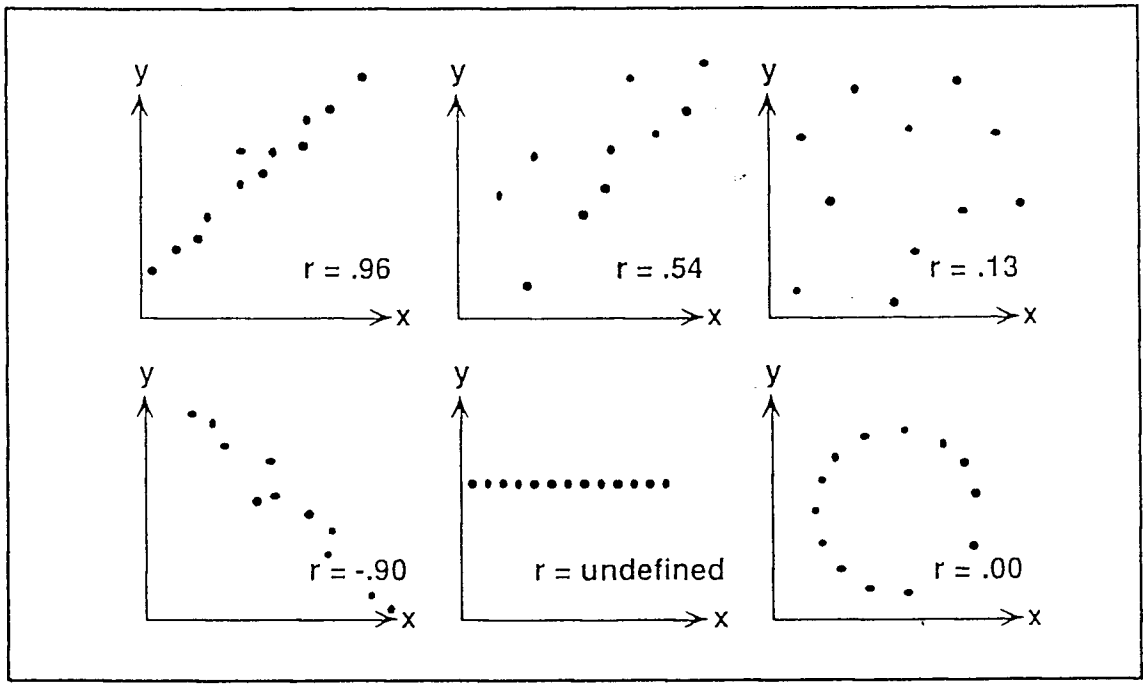
where s_x and s_y are the standard deviations of variables x and y.

The correlation coefficient is constrained between values of +1 and -1. A value of +1 is given by a perfect linear relationship ; -1 corresponds to a perfect inverse relationship; 0 means no linear relationship (see SIMPLE BIVARIATE PATTERNS...).
Crossplots, bivariate normal 90% contours , and correlation coefficients of Picaroon sandstone porosity correlations with other variables show interrelationships that match intuition (see PEARSON CORRELATION COEFFICIENTS...). At zero correlation, the bivariate normal distribution is isotropic and generates circular proportion contours. With increasing correlation, the contours become increasingly distended ellipses. For perfect correlation, the ellipses become lines with an orientation that reflects a positive or inverse relationship.

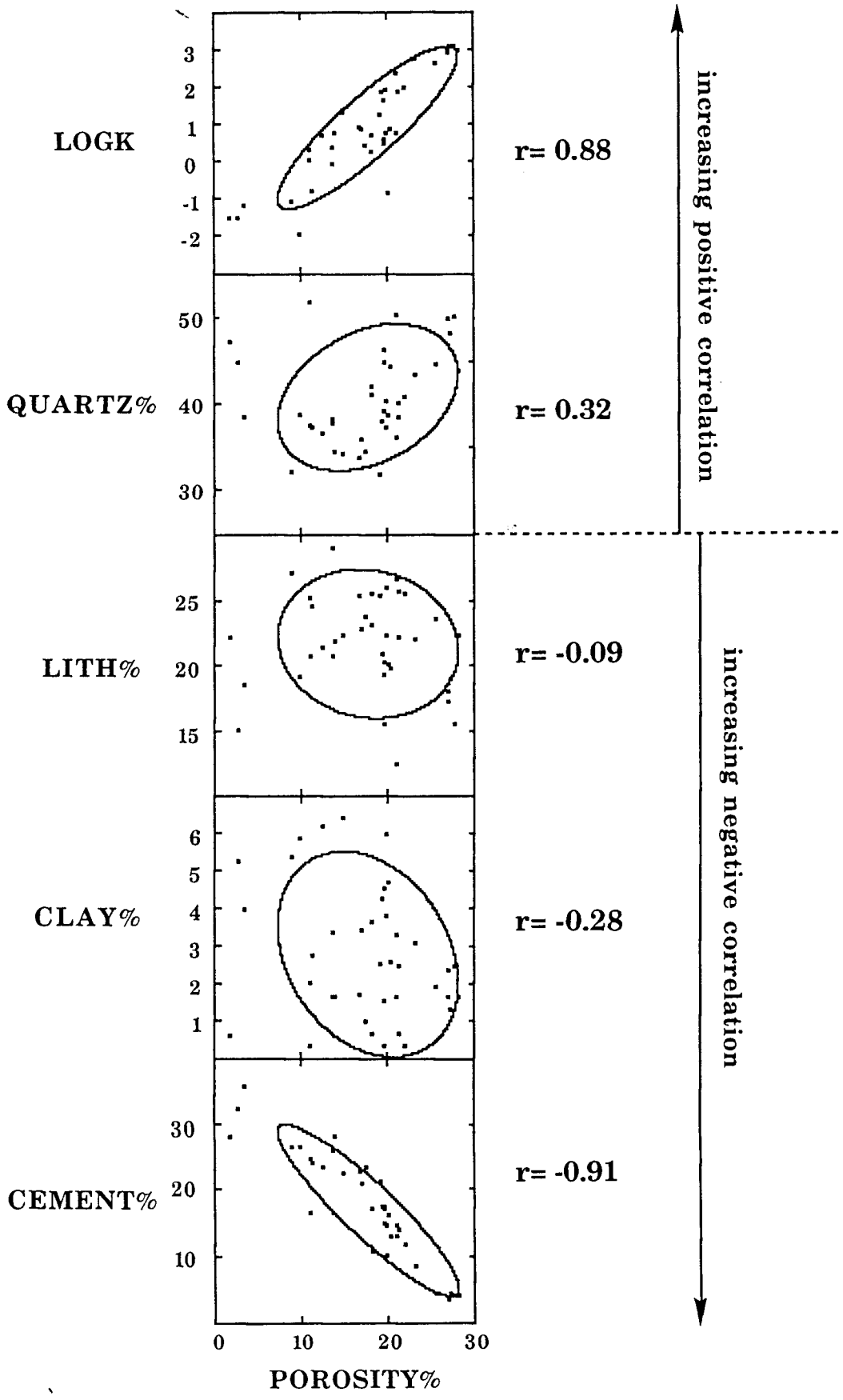
BIVARIATE NORMAL DISTRIBUTION CONTOURS BASED ON MEANS, VARIANCES, AND COVARIANCES FITTED TO (a) POROSITY-PERMEABILITY PLOT AND (b) POROSITY-LOGARITHMIC PERMEABILITY PLOT OF PICAROON SANDSTONES



CORRELATION COEFFICIENTS OF SOME SIMPLE BIVARIATE PATTERNS



PICAROON SANDSTONE PEARSON CORRELATION COEFFICIENTS OF POROSITY VERSUS LOG PERMEABILITY AND COMPOSITIONAL CONTENT



THE CORRELATION MATRIX

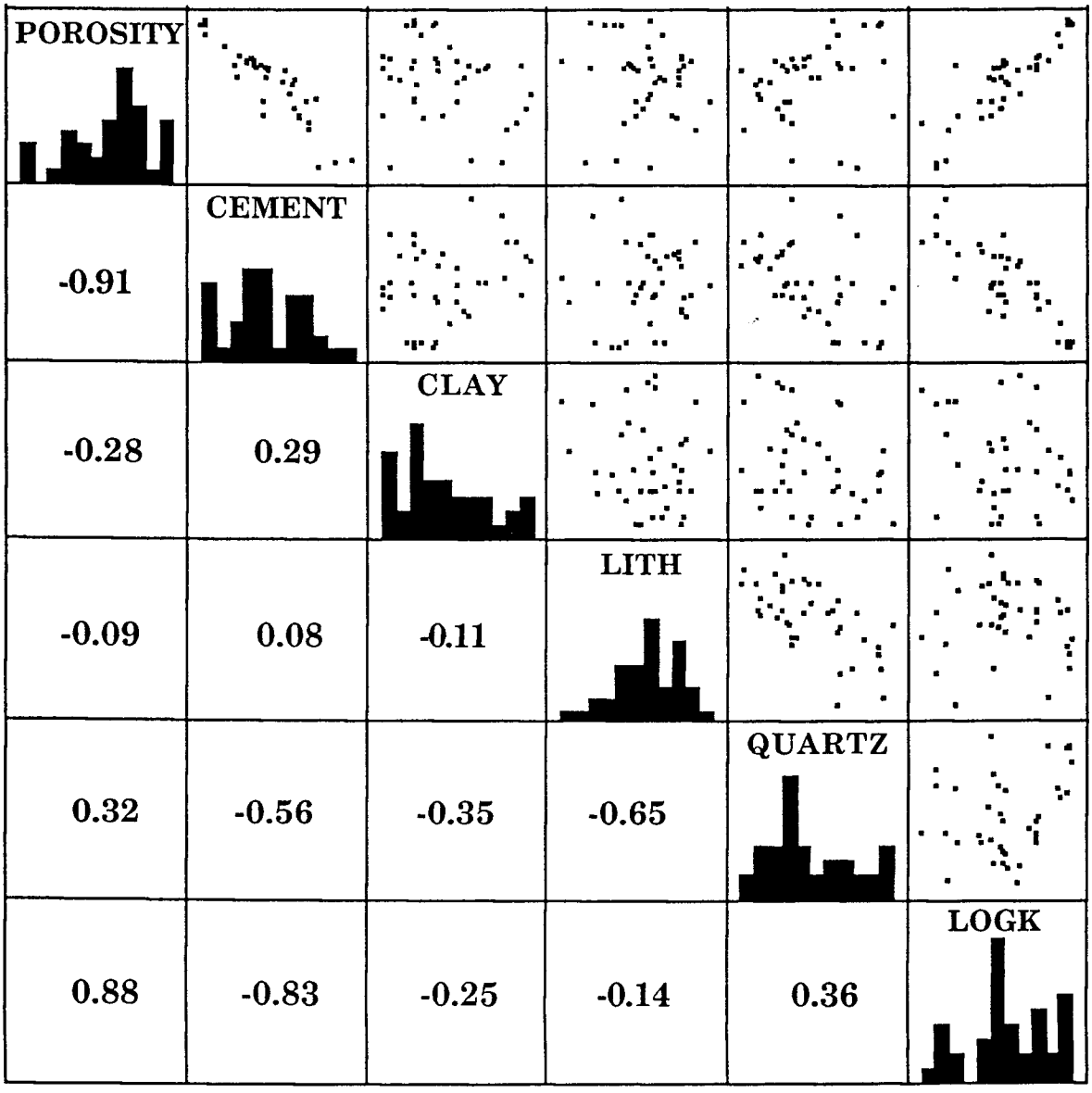
There will be many instances when correlation coefficients are calculated between all possible pairs of a number of variables. The results can be tabulated in the form of a correlation matrix, where the rows and columns are matched with the variables and used to assign pair assignations to each matrix cell. The correlation of variable x with y is the same as the correlation of variable y with x, so the matrix is symmetrical about the leading diagonal (upper left to lower right). The leading diagonal cells will all contain ones, because any variable is perfectly correlated with itself.

Scatterplots, histograms, and Pearson correlation coefficients are shown in the PICAROON SANDSTONE SCATTERPLOT MATRIX.. which has been designed to show the maximum amount of information, by restricting correlation coefficients to the lower part of the matrix (because values in the corresponding upper half cells are the same). The coefficients reflect the degree of linear trend that exists in the matching scatter plot.

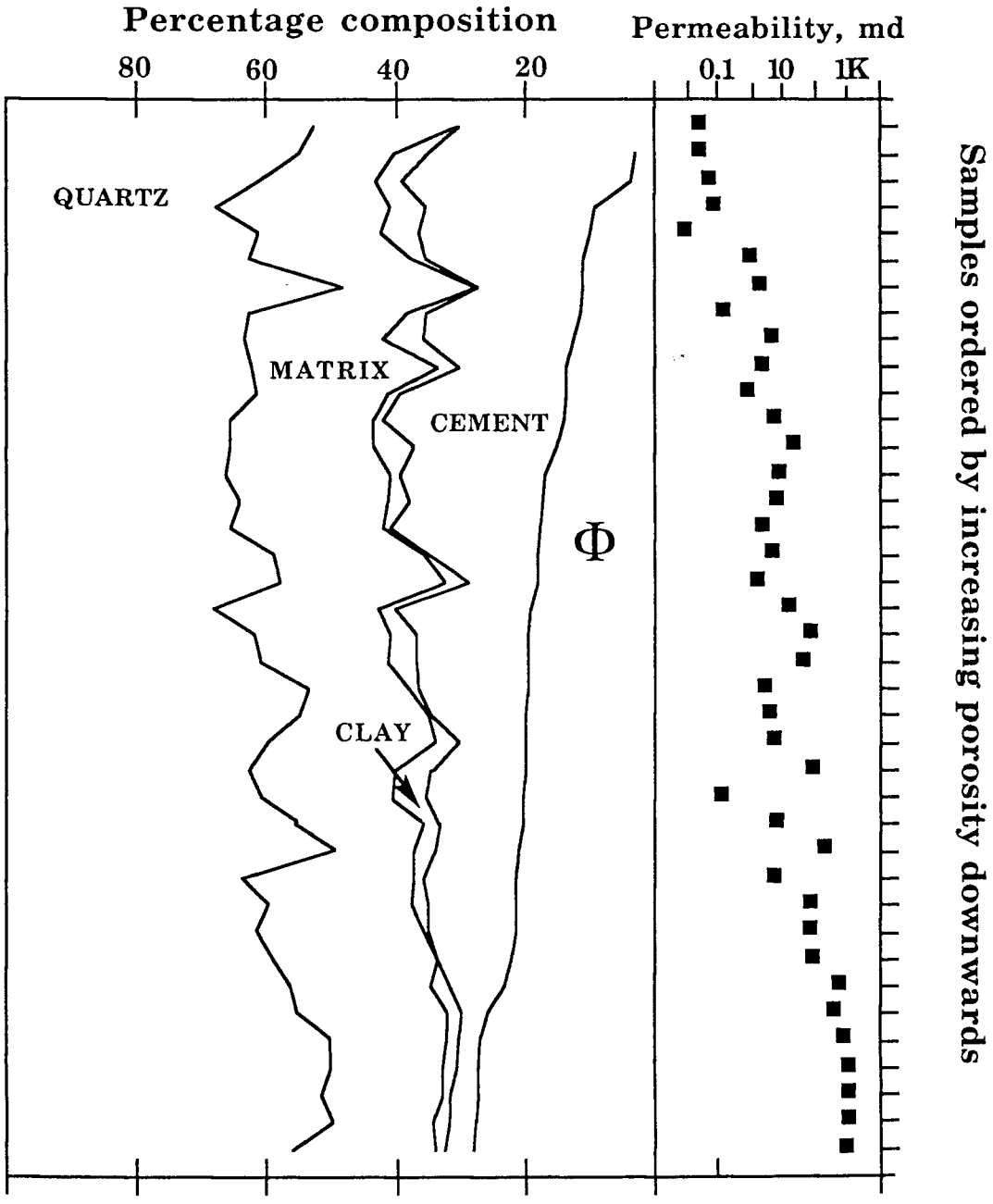
Particularly striking are the high positive (0.88) and high negative (-0.83) correlations between logarithmic permeability and porosity and cement. These associations are not surprising, but will be analyzed in more detail later. The highest correlation (-0.91) exists between porosity and cement -- in fact, an almost perfect inverse relationship whose petrographic and genetic significance was discussed by Taylor (1990). The almost perfect inverse relationship is brought out when the Picaroon sandstone compositional profile is reordered with respect to increasing porosity (see ORDERED BY POROSITY..).

There are some other interesting correlations among the compositional variables. However, their interpretation is fraught with major difficulties, because the variables collectively make up a closed-system. As a result, much of the correlation structure is erroneous. The values and even the signs (positive or negative) may be artificially induced distortions of their hypothetical true values if they had been enumerated in an open system. The problem and the results of a remedial correlation procedure are discussed under the next topic of "Closed Correlations".

PICAROON SANDSTONE SCATTERPLOT MATRIX AND PEARSON CORRELATIONS OF COMPOSITION AND LOGARITHMICALLY-SCALED PERMEABILITY



**ORDERED BY POROSITY:
COMPOSITION-PERMEABILITY PROFILE
OF THE PICAROON SANDSTONES**



CLOSED CORRELATIONS

Most conventional statistics texts do not consider the problem of false correlations that are induced by closure. Closure occurs when the analyzed variables sum to a fixed constant. When there are n variables in a closed system, there are only (n-1) independent variables. An entire dimension is lost. Algebraic (and geometrical) distortions will result when analysis is made of these variables as if they were a fully open system. The results are immediately obvious in the case of two or three closed variables (see CLOSURE...). The effects persist up to a surprisingly large number of closed variables. Sedimentary petrographers, igneous petrologists, and geochemists who choose to ignore this factor do so at their own peril. . Many apparent "patterns" in plots of closed data have been shown to be artifacts of closure.

Following extensive research Chayes (1971) correctly theorized that the solution to the closure problem lay through the analysis of ratios between proportions rather than the proportions themselves. He concluded that the best way to assess the meaning and significance of a correlation matrix was through comparison with a hypothetical closed array which had the same means and variances but calculated from a model in which the correlations of the variables would have been zero if they were in open form. Therefore, the intercorrelations seen on this comparison matrix would be caused purely by closure effects.

Unfortunately, Aitchison (1986) demonstrated that there are an infinite number of possible comparison matrices that will satisfy these conditions rather than any unique solution. In his book, Aitchison (1986) offered three procedures to analyze closed matrices, with a selection to be made, depending on what the reader intended to do with the results. A popular choice is the computation of a centered logratio covariance matrix, Γ, by the following method:

(1) If the compositional data are in percent, convert them to proportions;
(2) Divide each proportion by the geometric mean of that observation's compositional proportions and take its logarithm

$$\text{i.e. } \log\left(\frac{X_i}{g(X)}\right) \text{ which is equivalent to } \log(X_i) - \bar{\log}(X)$$

(3) Compute the covariance matrix of these centered log ratios, which is Γ, the centered logratio covariance matrix;
(4) Compute the standardized centered logratio matrix, which is the corrected correlation matrix that, hopefully, is free from closure effects.

A centered logratio correlation matrix was computed for the five compositional variables of the Picaroon sandstones (see GAMMA CORRELATIONS...). To do this, percentages were converted to proportions, then into centered logratios, and finally, these "open" variables were intercorrelated, and also correlated with logarithmic permeability.

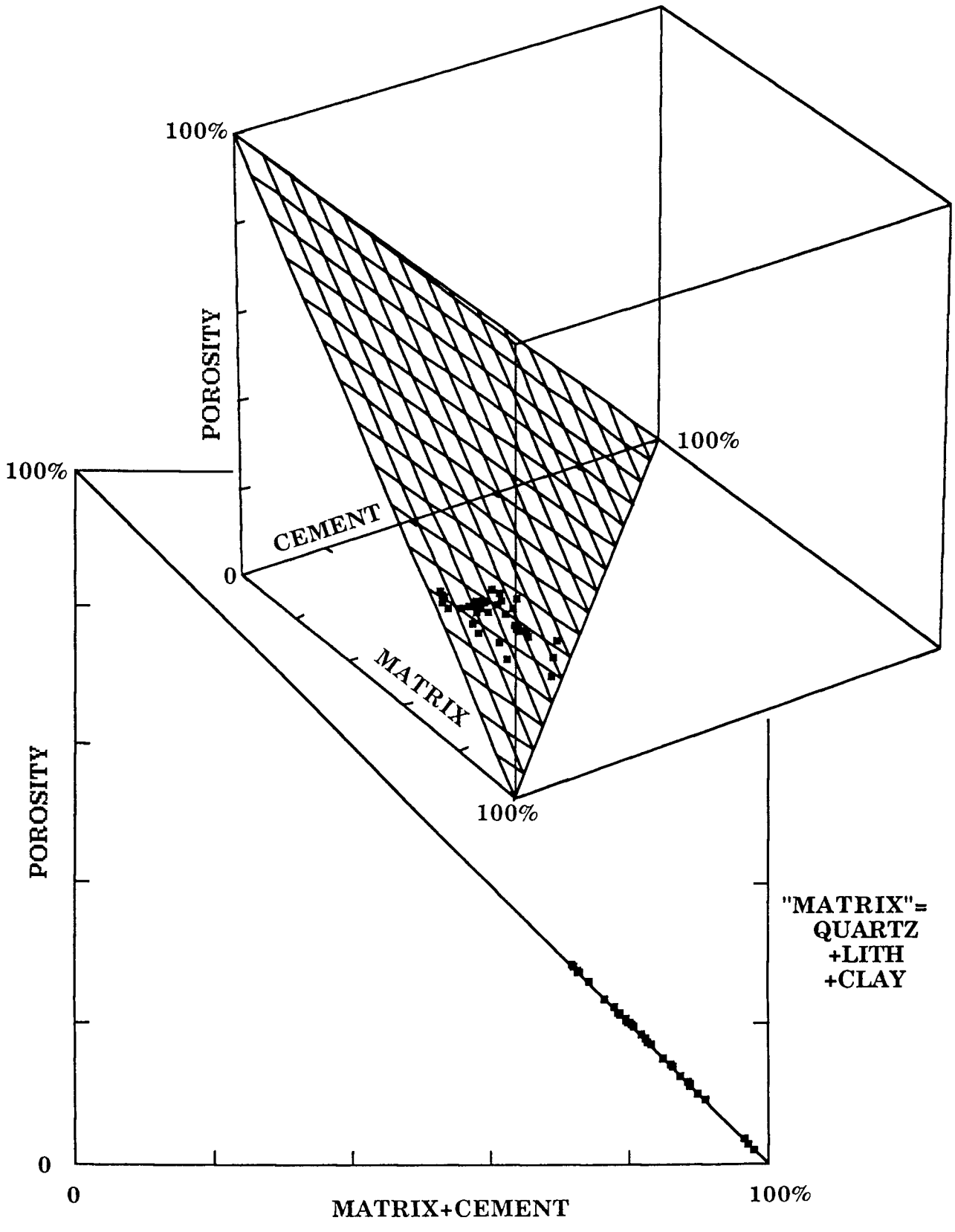
Collectively, the gamma correlation coefficients show both moderate similarities and striking differences with the correlations of their closed counterparts. Of particular interest are the intercorrelations between lithology fragments, quartz, and clay content. Are these statistically significant or do they represent little more than sample estimate variation about a population correlation coefficient of zero? This question can be addressed with a *t*-test, as described in the next section.

REFERENCES

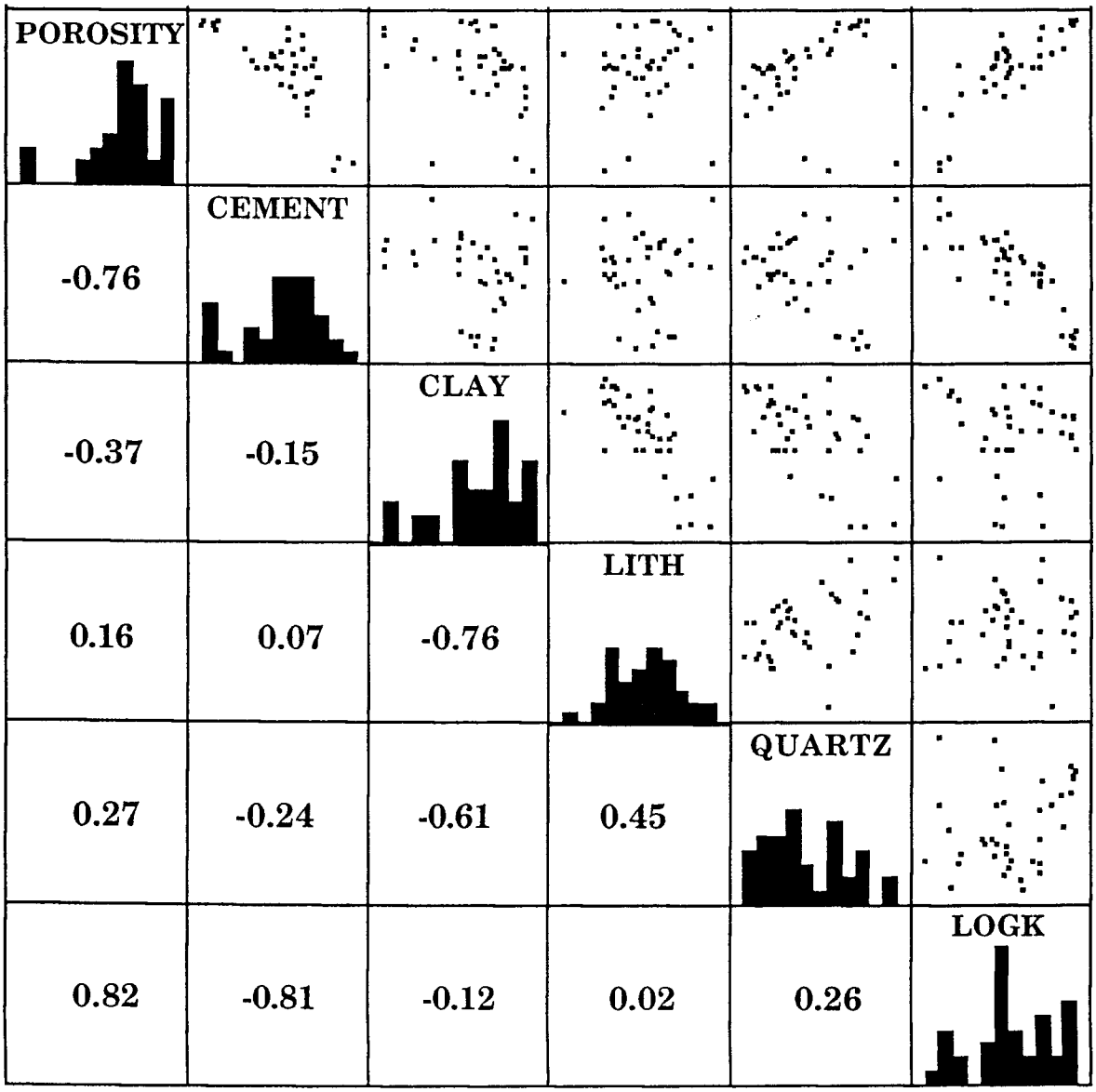
Aitchison, J., 1986, The Statistical Analysis of Compositional Data: Chapman and Hall, London, 416 pp.

Chayes, F., 1971, Ratio Correlation: Univ. of Chicago, Chicago, 99 pp.

CLOSURE IN COMPOSITION VARIABLES
ILLUSTRATED BY BIVARIATE AND TRIVARIATE
CONDENSATIONS OF PICAROON SANDSTONE DATA



PICAROON SANDSTONE SCATTERPLOT MATRIX
AND GAMMA CORRELATIONS OF LOG-RATIO COMPOSITIONS
AND LOGARITHMICALLY-SCALED PERMEABILITY



SIGNIFICANCE OF CORRELATION

The Pearson product-moment correlation coefficient, r is calculated for a sample and is an estimate of the population parameter, ρ. Because the values of r are constrained between the limits of +1 and -1, the sampling distribution of r is highly skewed near the limits of this range. When ρ= 0, the distribution of r is symmetrical, although not exactly normal. The null hypothesis of no correlation:

$$H_0: \rho = 0$$

can be tested using a *t*-distribution, where:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with } (n\text{-}2) \text{ degrees of freedom.}$$

If the calculated value for *t* exceeds the tabulated critical value for a two-tailed test at a selected significance level, then the null hypothesis of no correlation is rejected.

In the case of the Picaroon sandstone data, the sample size, n= 39, so that the number of degrees of freedom, ν= 37. Then:

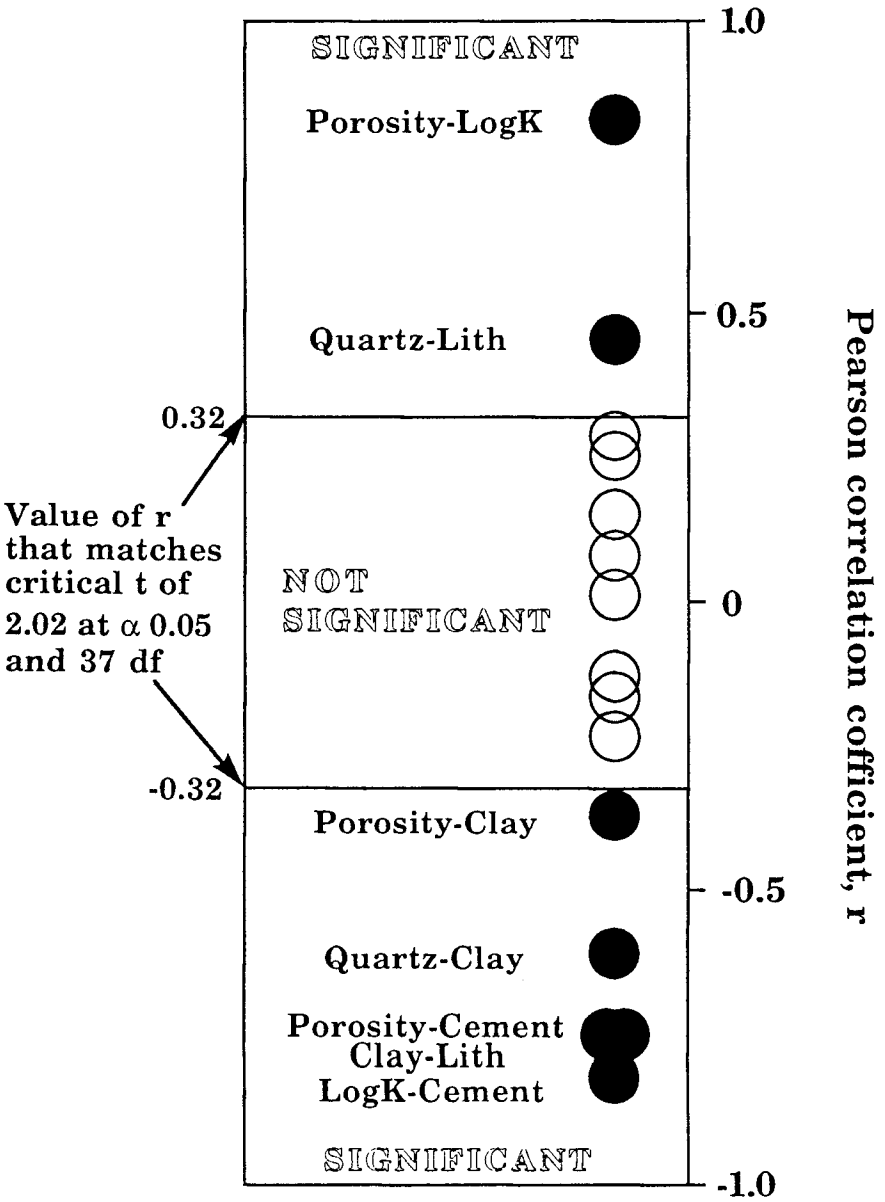
Critical t value @ a 0.05 and 37 df = 2.02

Substituting this value into the formula for *t* above, gives the critical absolute value of r= 0.32, which must be exceeded by a Picaroon sandstone sample estimate, in order to be considered to be significantly different from an expectation of zero.

The critical value can be used to discriminate potentially significant correlations in the Picaroon sandstone gamma correlation matrix (see POTENTIALLY SIGNIFICANT CORRELATIONS ...). This method of scanning correlation matrices is commonly done as a data exploration tool. However, we should recognize the expectation that we will falsely reject the null hypothesis by a proportion equal to the chosen significance level. In other words, the acceptance of a given risk level will result in a proportionally small proportion of significant correlations that are spurious.

The strongest associations seem to be between porosity and log permeability and their inverse relationships with cement. At a lower, but significant level, is the positive association between quartz and lithology fragments, and their common negative correlations with clay. This pattern probably reflects a grain size control of composition. A weaker and barely significant negative correlation of porosity with clay suggests a tendency for decrease in porosity with finer grain size.

POTENTIALLY SIGNIFICANT CORRELATIONS IN THE PICAROON SANDSTONES OF LOG RATIO COMPOSITIONS AND LOGARITHMICALLY -SCALED PERMEABILITY DISCRIMINATED THROUGH USE OF A t-TEST
(Significant correlations: labelled closed circles; correlations considered not to be significant: unlabelled open circles.)



RANK CORRELATION

Because it is impossible to estimate parameters measured on an ordinal scale (ordered categories), some form of non-parametric measure of association must be used for this type of data. If measurements of two ordinal variables are ranked in order for each of two variables, a Spearman's rank correlation coefficient, r', may be calculated from:

$$r' = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

where n is the number of individuals in the sample, and D_i is the difference between the ranks of the ith individual measured on the two variables. Where ties occur, the ranks are averaged. The statistic is based on the $n!$ possible different combinations of the ranks of the two variables. However, the mathematics are such that the Spearman coefficient will have the same value as the Pearson coefficient computed from the ranks, in instances where no tied ranks occur.

The possible values are constrained between +1 and -1, with the same implications as the Pearson correlation coefficient. The potential significance of the computed r' is also examined in a similar manner by an approximate t-test of the hypothesis that ρ' is zero, based on:

$$t = \frac{r'\sqrt{n-2}}{\sqrt{1-r'^2}}$$

The rank correlation coefficient is also useful for continuously scaled data in cases where the variables are drastically different properties with radically different units. Because the Pearson correlation is a measure of **linearity** between variables, a strong non-linear relationship may give a disappointingly weak Pearson coefficient. However, by substituting a Spearman measure for ranked data, a coefficient is selected that is sensitive to a **monotonic** pattern in the data, regardless of whether the trend is non-linear.

As an example of the utility of the Spearman rank correlation coefficient, we shall consider the potential relationship between internal surface area and porosity in the Picaroon sandstone samples. Now, one of the more common versions of the Kozeny-Carman equation takes the form of:

$$k = \frac{A\Phi^3}{S_0^2(1-\Phi)^2}$$

which relates permeability, k, to porosity, Φ, and specific surface area, S_0, and where A is a constant. We can compute a generalized measure of specific surface area, S, from the equation:

$$S = \sqrt{\frac{\Phi^3}{k(1-\Phi)^2}}$$

The values for S are listed in the table COMPUTATION OF SPEARMAN... The table shows the results of ranking the Picaroon sandstones, first with respect to porosity, then with respect to specific surface area, and the computation steps that led to the calculation of the Spearman rank correlation coefficient, r' of -0.58. The table also shows the result of a t-test of the hypothesis of null correlation which was rejected in favor of accepting a significant negative association between specific surface area and porosity.

The power of the Spearman test and its meaning for this example are shown well by the comparative crossplots of porosity versus specific surface area (see POROSITY AND SPECIFIC SURFACE AREA...) The plot of the raw variation shows no immediate trend and the Pearson correlation coefficient of -0.23 indicates a weak negative linear association that is not significant. However, when replotted in rank form the tendency for monotonic decrease in specific surface area with porosity becomes much more obvious, and the Spearman correlation coefficient picks up a significant negative trend. The physical interpretation of the pattern is that pore sizes tend to be larger at higher porosities. In addition, the greater spread at lower porosities suggests a range of pore sizes, in contrast with higher porosities, where the pore size seems to be more uniform.

PICAROON SANDSTONE: COMPUTATION OF SPEARMAN RANK CORRELATION COEFFICIENT BETWEEN POROSITY AND SPECIFIC SURFACE AREA

POROSITY %	RANKED	SPECIFIC SURFACE AREA S	RANKED	D^2
19.6	21.5	0.016	15	42.25
23.3	33	0.006	4	841.00
13.8	11	0.064	34	529.00
18.2	17.5	0.043	27	90.25
17.6	16	0.056	31	225.00
19.6	21.5	0.063	33	132.25
19.8	23	0.056	32	81.00
21.2	29	0.053	30	1.00
22.0	32	0.014	11	441.00
21.4	30.5	0.015	12.5	324.00
21.4	30.5	0.015	12.5	324.00
18.2	17.5	0.071	35	306.25
20.0	24.5	0.049	29	20.25
20.4	27	0.044	28	1.00
12.5	9	0.024	17	64.00
15.0	13	0.015	14	1.00
16.8	14	0.030	23	81.00
17.2	15	0.032	24	81.00
19.3	19	0.025	18	1.00
11.1	6	0.042	26	400.00
14.1	12	0.027	19	49.00
10.0	5	0.351	39	1156.00
19.5	20	0.013	10	100.00
20.0	24.5	0.012	9	240.25
11.5	8	0.114	37	841.00
13.7	10	0.038	25	225.00
9.00	4	0.105	36	1024.00
3.50	3	0.028	20	289.00
20.3	26	0.318	38	144.00
2.80	2	0.028	21	361.00
2.00	1	0.017	16	225.00
11.2	7	0.029	22	225.00
21.1	28	0.008	7	441.00
28.2	39	0.007	6	1089.00
27.8	38	0.006	3	1225.00
27.4	37	0.006	1	1296.00
27.2	36	0.006	2	1156.00
25.7	34	0.009	8	676.00
27.1	35	0.007	5	900.00

$$\sum D^2 = 15648.5$$

$$r' = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

$$r' = -0.58$$

$$t = \frac{r'\sqrt{n-2}}{\sqrt{1-r'^2}}$$

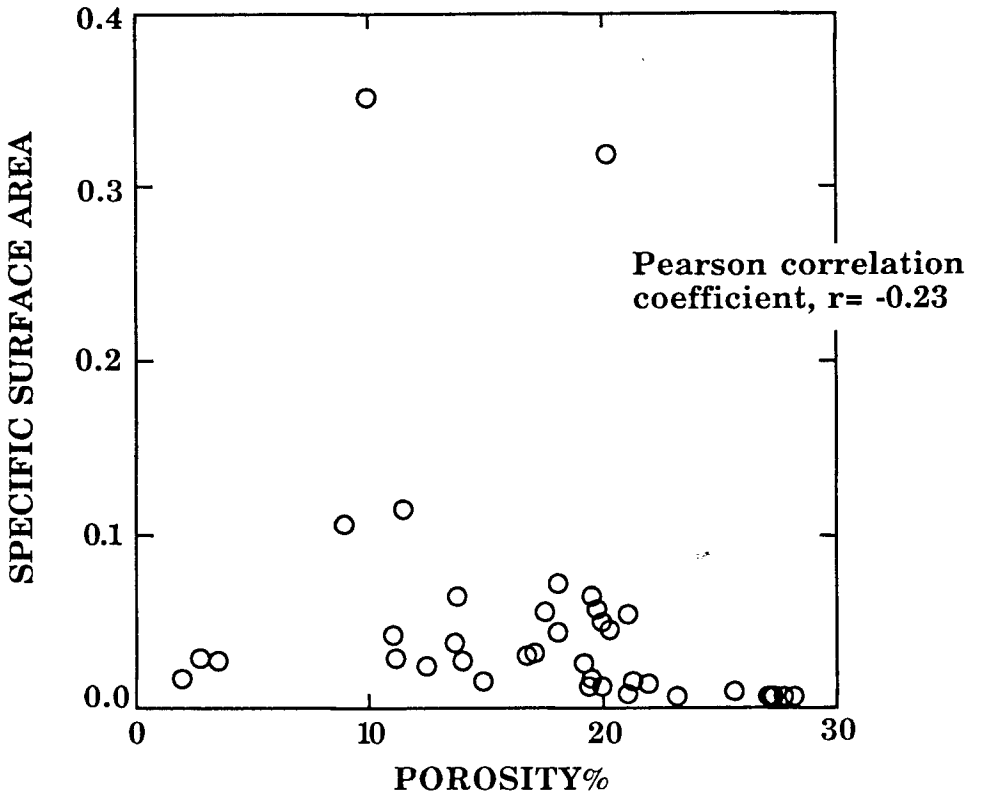
$$t = 4.38$$

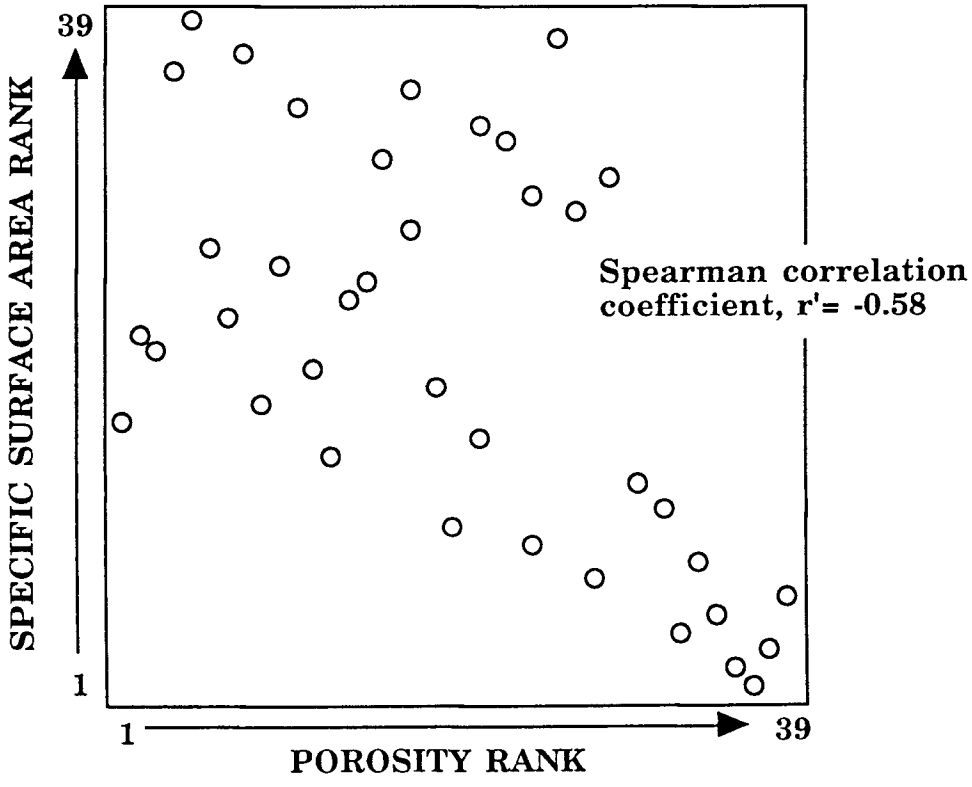
n= 39
df= (n-2)= 37

Two-tailed test
critical t = 2.02
@ 0.05a and 37 df

The null hypothesis is rejected and it is accepted that there is a significant negative association between specific surface area and porosity.

PICAROON SANDSTONE:
PLOTS OF POROSITY AND SPECIFIC SURFACE AREA
(a) AS RAW VARIABLES, (b) AS RANKINGS





PRELIMINARY IDEAS ON LINEAR REGRESSION

Measures of correlation are an expression of the intensity of an association between variables. In the case of the Pearson product-moment correlation coefficient, the statistic gauges the degree of linear trend. The purpose of regression analysis is to isolate some functional trend that relates changes in one variable with changes in another. The function can then be used for purposes of prediction and statements can be made with regard to the likely magnitude of error. Regression analysis is based on the principle of least squares, with squared deviations from the trend attributed to error and having a normal distribution about the trend.

Simple linear regression relates the variation of one variable, y, with respect to another, x, in terms of a linear function:

$$Y = a_0 + a_1 X + e$$

where X is called the independent variable (which is assumed to have no error) and Y is the dependent variable with an associated random error, e. The equation of the line itself is:

$$\hat{Y} = a_0 + a_1 X$$

where the "hat" on Y signifies that it is the prediction of Y, given a value of X. The quantities a_0 and a_1 are unknown constants whose values will be solved by regression analysis. They represent the intercept and slope of the line, respectively.

To clarify some of these ideas, we will examine a regression analysis of the logarithmic permeability on porosity in the Picaroon sandstone sample. We have already seen that there is a high, positive and significant Pearson correlation between these two variables. If we equate log permeability with Y and porosity with X, then we will be making predictions of log permeability, as the dependent variable, based on given values of porosity as the independent variable. An extensive geological and engineering literature has been published on precisely this problem, because of the great economic benefits of a prediction equation that performs well. Permeability is a crucial control of reservoir producibility but measurements are generally infrequent and mostly limited to core analyses. Wireline logs are widely available but do not provide direct measurements of permeability. However, if a strong relationship could be established with porosity, then porosity logs could be transformed into profiles of permeability. (In real life, additional factors, such as non-linearities, other controlling variables, etc. will complicate the situation, but regression analysis can be expanded to incorporate them).

The line from the computation of the regression of log permeability on porosity in the Picaroon sandstones shows a good visual fit to the data (see LINEAR REGRESSION...). However, a comparison between predicted and actual values

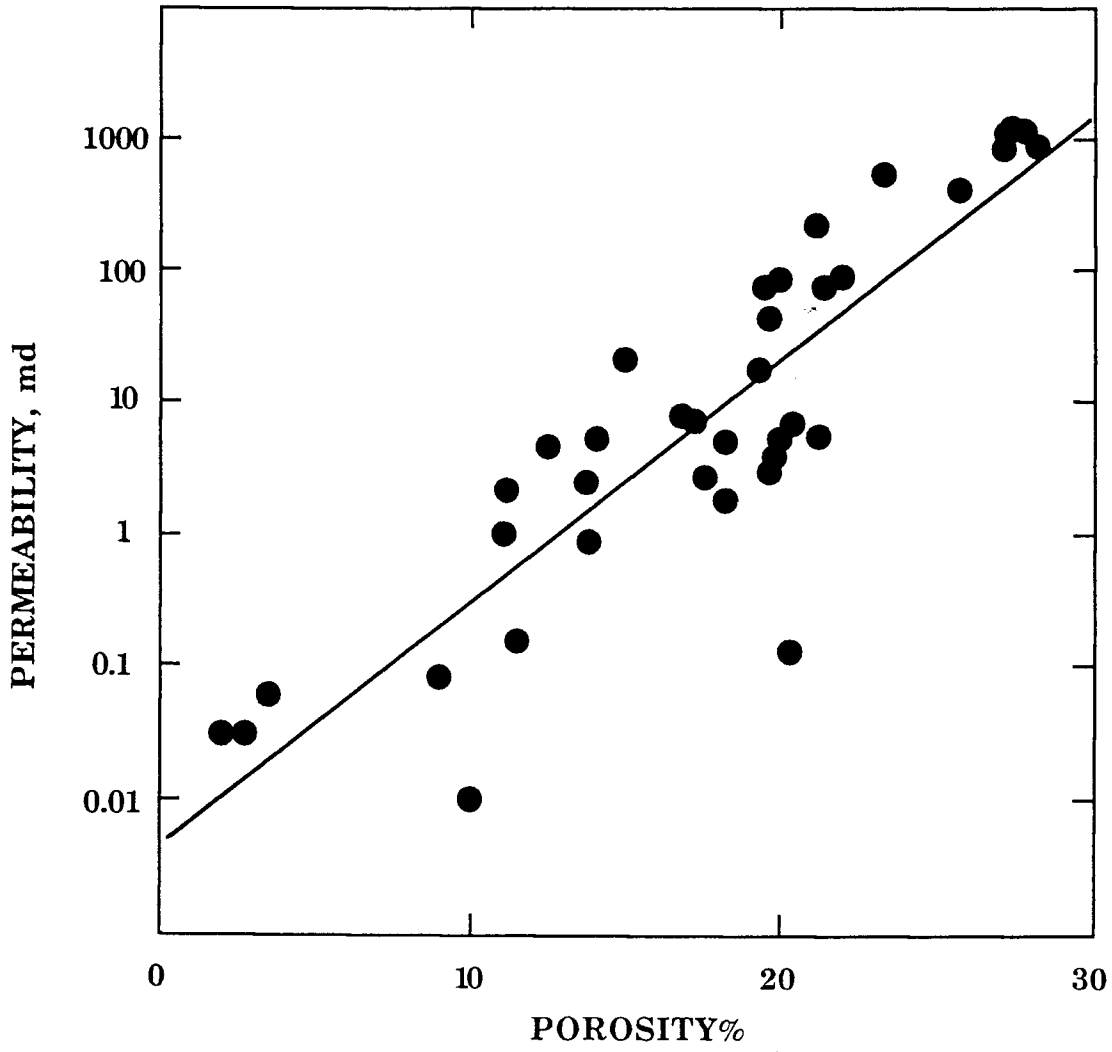
of permeability on a depth profile has both good and bad features (see PERMEABILITY DEPTH PLOT...). At first glance, the overall match looks quite good. However, closer inspection shows that the regression prediction tends to overestimate the extreme lows and underestimate the extreme highs. If anything, the function appears to produce an excellent estimation of a moving average, rather than reproducing the extremes. This is indeed the case, because the regression estimates the average value of Y given any value of X. This is the optimal choice, because it generates the minimum potential squared error when the prediction is compared with the actual value (see ARTIFICIAL DATA...). This fundamental regression property of estimating the mean is discussed at length in both general statistical texts and papers that focus on permeability prediction. So, for example, Wendt et al (1986) (p.205) pointed out that not only were the extremes under- and overestimated but that the logarithmic scale of permeabilities exacerbated the problem at the high permeability end. As a remedial measure, they advocated preferential weighting of high and low values in order to pivot the line and honor the extremes more closely. To some degree, the analysis strategy will depend on the motives of the investigator. If the characterization of high permeability streaks is important, then remedial steps may be called for. If the permeabilities will be coarsely averaged to provide statistics for reservoir simulation models, then no modifications may be necessary because the regression estimates mean values already.

REFERENCES

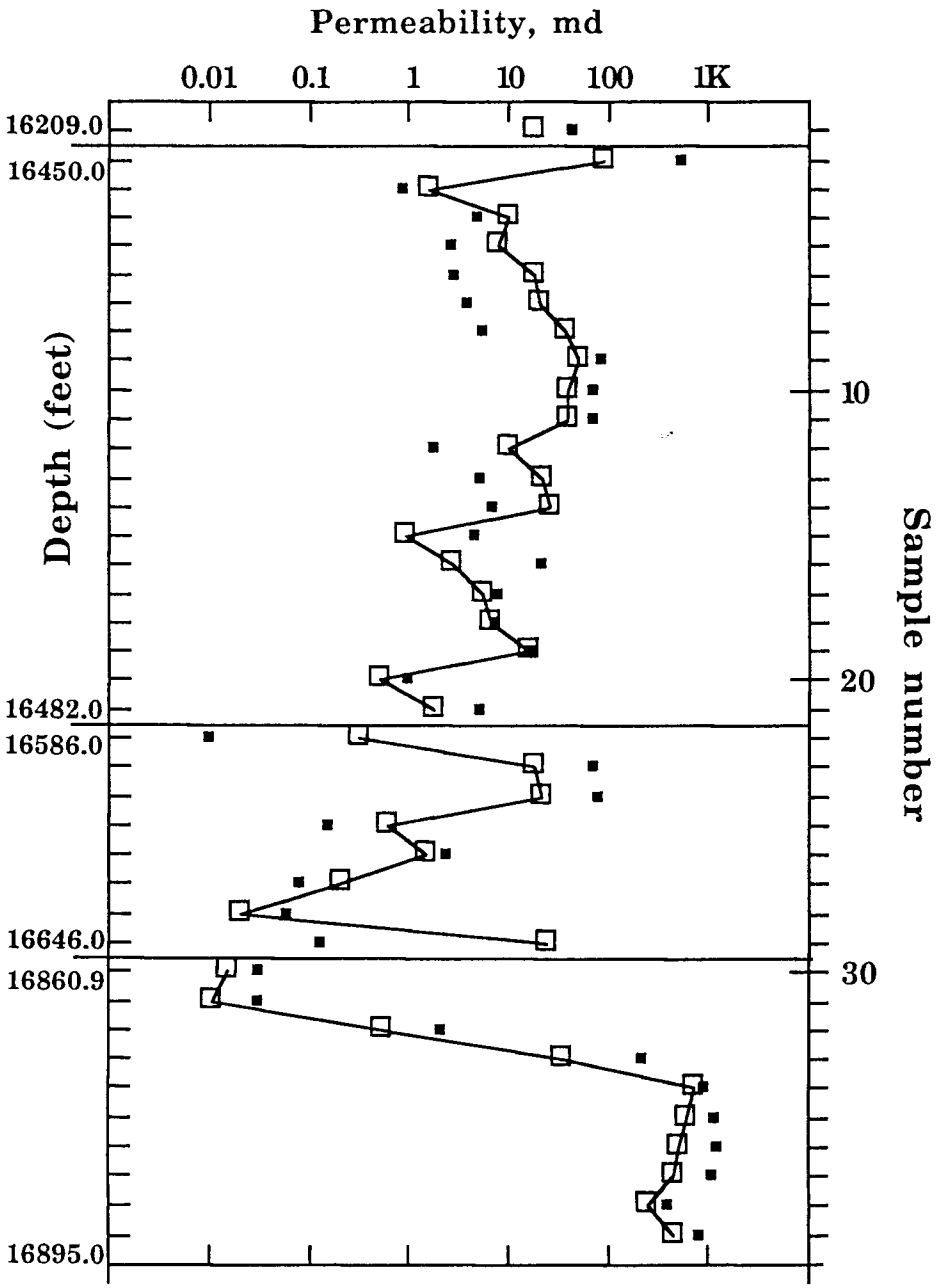
Campbell, D.T., and Stanley, J.C., 1966, Experimental and quasi-experimental designs for research: Rand McNally, 78 pp.

Wendt, W.A., Sakurai, S., and Nelson, P.H., 1987, Permeability prediction from well logs using multiple regression: in Reservoir Characterization (eds. Lake and Carroll), Academic Press, p. 181-221.

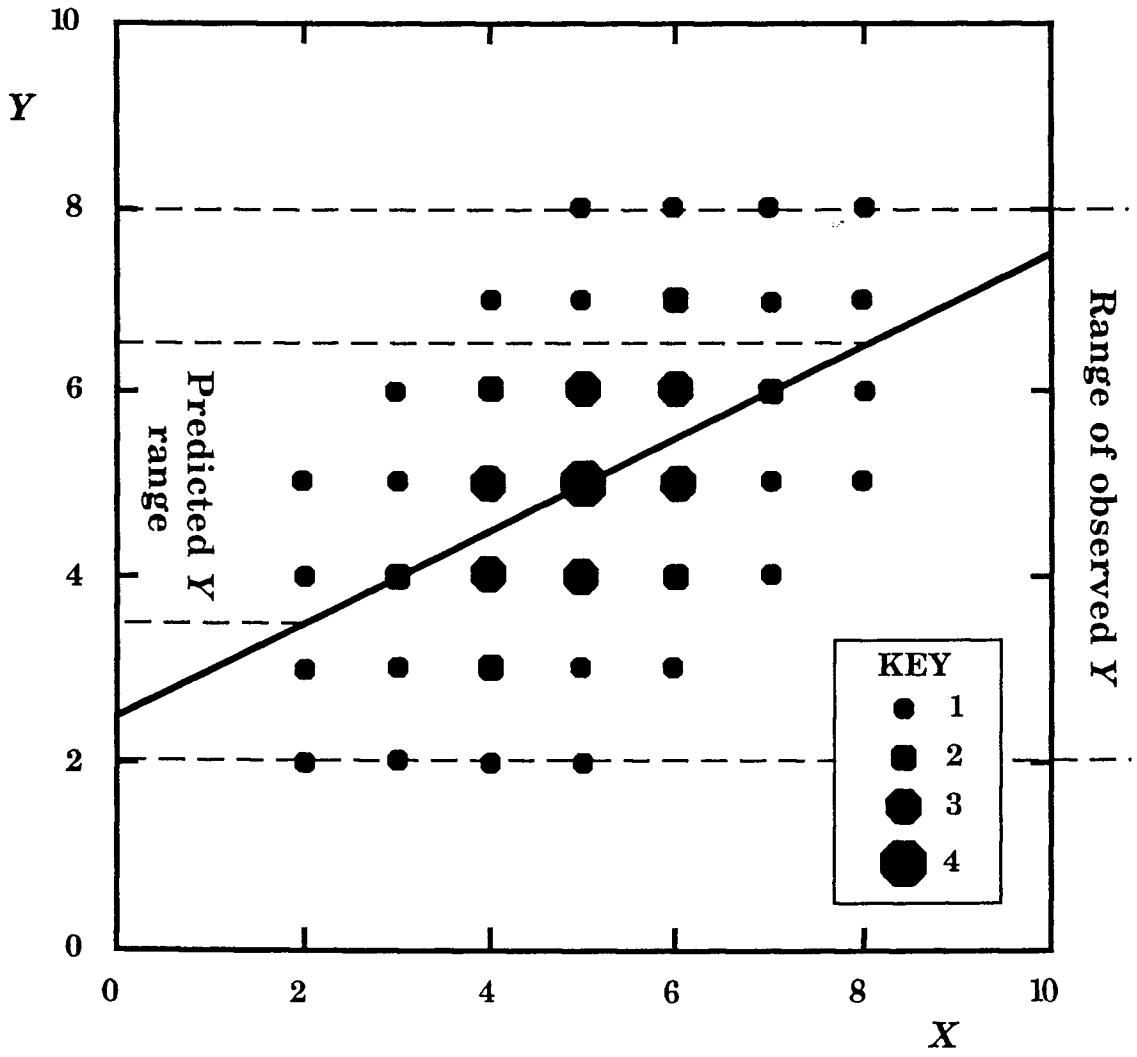
LINEAR REGRESSION OF LOG PERMEABILITY ON POROSITY IN PICAROON SANDSTONES



PERMEABILITY DEPTH PLOT OF ESTIMATES FROM LINEAR REGRESSION ON POROSITY (OPEN SQUARES) AND RAW PERMEABILITIES (CLOSED SQUARES) IN PICAROON SANDSTONES



REGRESSION OF Y ON X FOR ARTIFICIAL DATA, WHICH SHOWS THE EFFECTS ASSOCIATED WITH THE BEST PREDICTION OF Y ON AVERAGE FOR ANY GIVEN X VALUE , INCLUDING THE CONTRACTION OF Y PREDICTION RANGE COMPARED WITH OBSERVED VALUES. THE SIZE OF THE SYMBOLS REPRESENTS THE NUMBER OF POINTS AT EACH LOCATION AS DEFINED IN THE KEY. DATA FROM CAMPBELL AND STANLEY (1966).



THE GEOMETRY AND CALCULUS SOLUTION OF THE REGRESSION LINE, AND ITS STATISTICAL ASSESSMENT

When a regression line of Y on X is fitted to a sample of bivariate data, the line is located such that the sum of the squared deviations of Y from the line is the minimum possible (see LINEAR REGRESSION OF Y ON X). The position and orientation of this line is determined by its intercept and slope. The estimates of these parameters can be solved uniquely from simultaneous equations developed from simple calculus.

Once the coefficients of the regression line have been estimated, some determination of its significance should be made. If there is no relationship between the variables x and y, then for an infinite population, the regression line will be horizontal (with zero slope) and an intercept on the Y axis equal to the mean value of Y. This regression line is a perfectly rational solution. If there is no relationship, then the value of X is immaterial and the best estimate is the mean value of Y. This estimate is the best, because the squared deviations are the minimum possible. So, the errors between predictions and actual values will be minimized.

Analysis of variance (ANOVA) can be used to assess whether a regression relationship accounts for a significant trend over and beyond the use of the mean value of Y (see SOURCES OF VARIATION...). The total sums of squares is subdivided between the sums of squares picked up by the regression and the sums of squares left over in the deviations about the line (or residuals).

The goodness-of-fit is the proportion of the total variation absorbed by the regression:

$$R^2 = \frac{SS_R}{SS_T}$$

This "coefficient of determination" is the square of the Pearson correlation coefficient between x and y.

An ANOVA table (see SOURCES OF VARIATION) reports the budget of the total sums of squares between regression and deviations, the number of degrees of freedom associated with each source, the mean square value (sums of squares divided by the degrees of freedom). An F-test of the value:

$$F = \frac{MS_R}{MS_D}$$

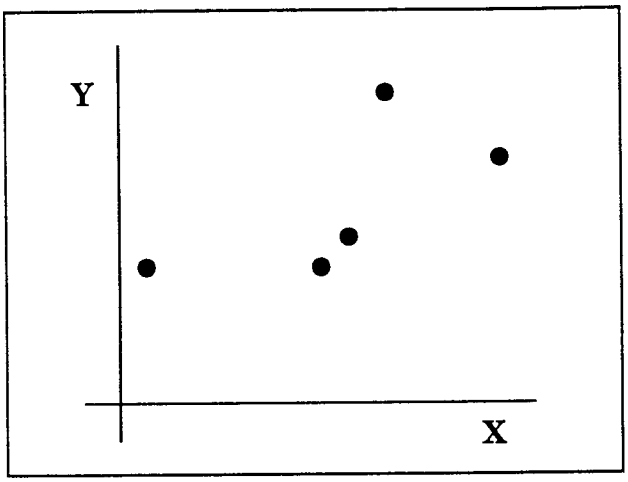
is used to test the null hypothesis that the sample estimate slope of the regression line is not significantly different from zero. If the calculated F-value exceeds the critical F-test value at 1 and (n-2) degrees of freedom and the selected significance level, the alternative hypothesis is accepted: the regression line does represent a significant trend.

LINEAR REGRESSION OF Y ON X

DATA SET

n observations

X_1	Y_1
X_2	Y_2
X_3	Y_3
.....	
X_n	Y_n



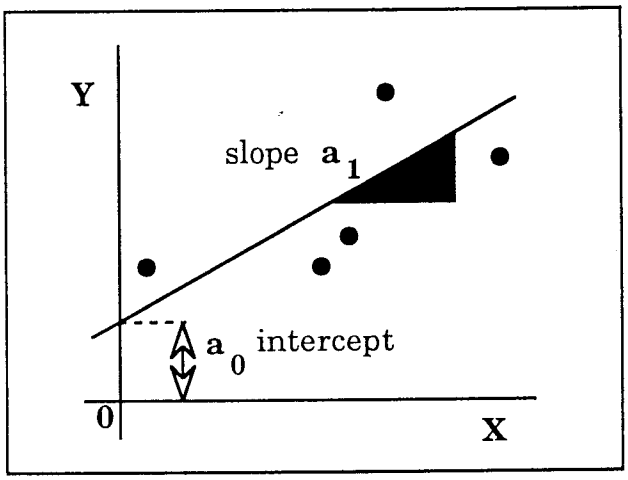
dependent (predicted) variable

independent (predictor) variable

$$\hat{Y} = a_0 + a_1 X$$

intercept

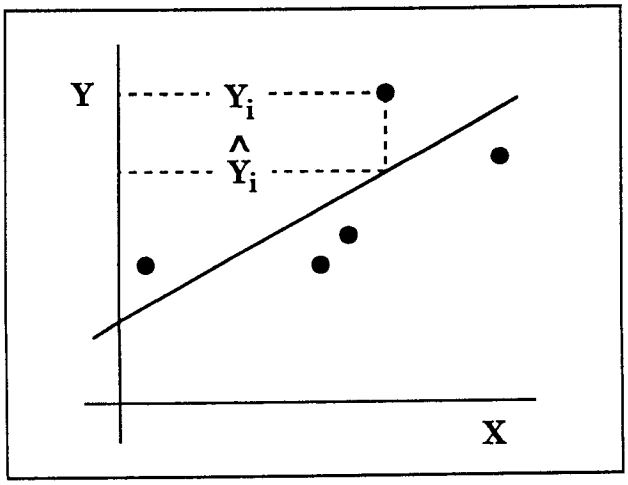
slope



The regression line of Y on X is fitted using the "principle of least squares", which minimizes the sum of the squared deviations of Y from its predicted value, $\hat{Y}$

$$\sum (Y_i - \hat{Y}_i)^2 = G$$

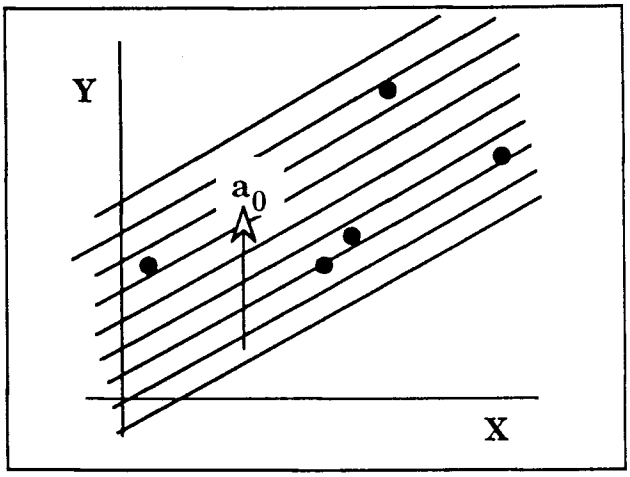
where G is the minimum possible value.

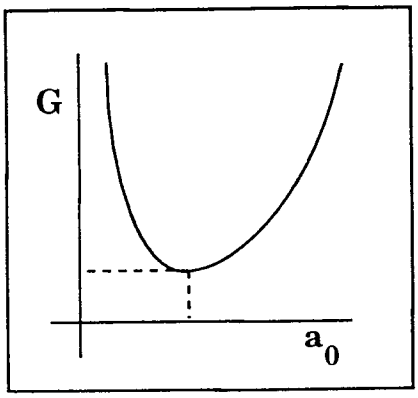


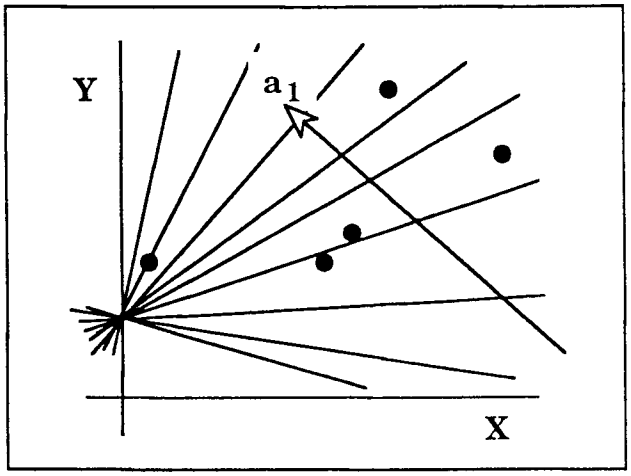
$$\sum(Y_i - \hat{Y}_i)^2 = G = \text{minimum}$$

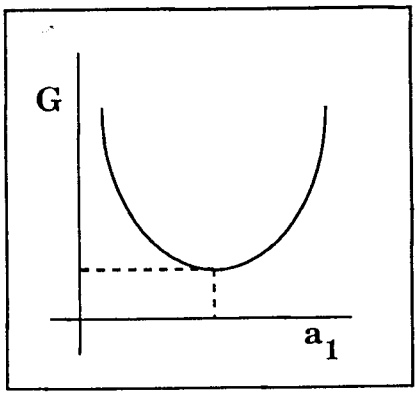
But... $\hat{Y}_i = a_0 + a_1 X_i$

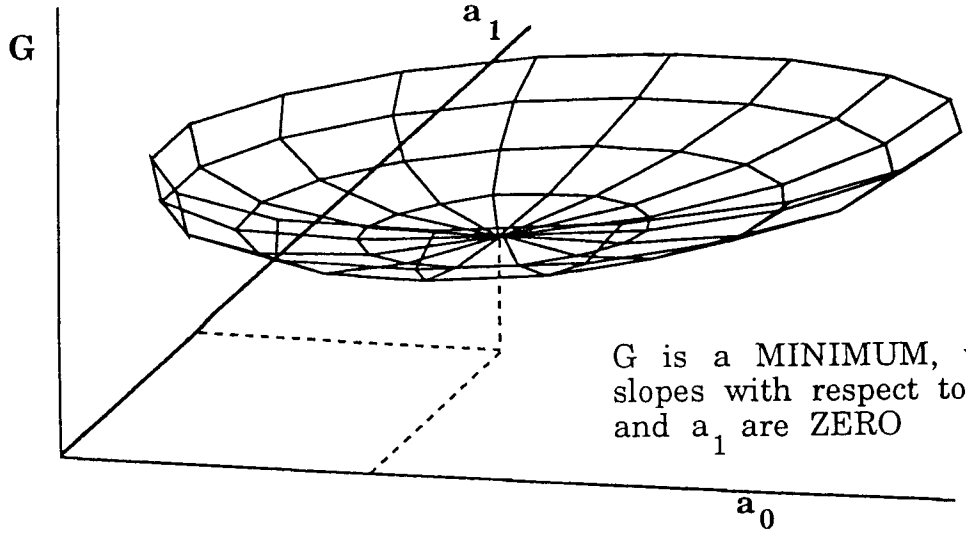
So... $\sum(Y_i - a_0 - a_1 X_i)^2 = G = \text{minimum}$











G is a MINIMUM, when slopes with respect to a_0 and a_1 are ZERO

The slope of an equation is given by the first differential i.e. for equation $y = f(x)$, the slope is dy / dx

If $\sum(Y_i - a_0 - a_1X)^2 = G$, G is a minimum when the partial differentials with respect to both a_0 and a_1 are zero i.e.

$$\frac{\partial G}{\partial a_0} = 0 \quad \text{and} \quad \frac{\partial G}{\partial a_1} = 0$$

Differentiating :

$$\partial G/ \partial a_0 = \sum - (Y_i - a_0 - a_1X_i) = 0$$
$$\partial G/ \partial a_1 = \sum - X_i(Y_i - a_0 - a_1X_i) = 0$$

Rearranging :

$$n a_0 + a_1 \sum X_i = \sum Y_i$$
$$a_0 \sum X_i + a_1 \sum X_i^2 = \sum X_i Y_i$$

Rewriting in matrix form :

$$\begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix}$$

$$\mathbf{X} \qquad \mathbf{A} \qquad \mathbf{Y}$$

$$\mathbf{XA = Y}$$

$$\mathbf{A = X^{-1} Y}$$

The vector A is the solution for the intercept a_0 and the slope a_1 of the regression line of Y on X

SOURCES OF VARIATION -- IS THE REGRESSION TREND OF Y ON X SIGNIFICANT?

Sum of squares, regression : $SS_R = \sum (\hat{Y} - \bar{Y})^2$

Sum of squares, deviation : $SS_D = \sum (Y - \hat{Y})^2$

Sum of squares, total : $SS_T = \sum (Y - \bar{Y})^2$

$$SS_T = SS_R + SS_D$$

GOODNESS - OF - FIT is the proportion of the total variation accounted for by the regression :

$$R^2 = SS_R / SS_T$$

R is equal to the correlation coefficient between X and Y.

ANALYSIS OF VARIANCE

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		

If this value exceeds the critical F-test value at 1 and (n-2) degrees of freedom at a preselected level of significance, then the null hypothesis that the variance about the trend is no different than the variance about the mean is rejected. In this case, the alternative hypothesis is accepted and the trend considered to be significant.

ALTERNATIVE REGRESSIONS: Y on X and X on Y ; THE REDUCED MAJOR AXIS (RMA)

For any set of bivariate data, x and y, two alternative regression lines may be computed: Y on X or X on Y. One can either predict Y **given** a value of X, or one can predict a value of X **given** a value of Y. The two alternatives attribute all the error to Y (Y on X) or all the error to Y. The regression lines intersect at the coordinates of the bivariate means (see COMPARISON OF REGRESSION...). The cosine of the angle between the two lines is equal to the Pearson correlation coefficient, r. For perfect correlation, the two lines coincide. when there is no correlation, the lines are horizontal and vertical axes locked onto the mean values of X and Y.

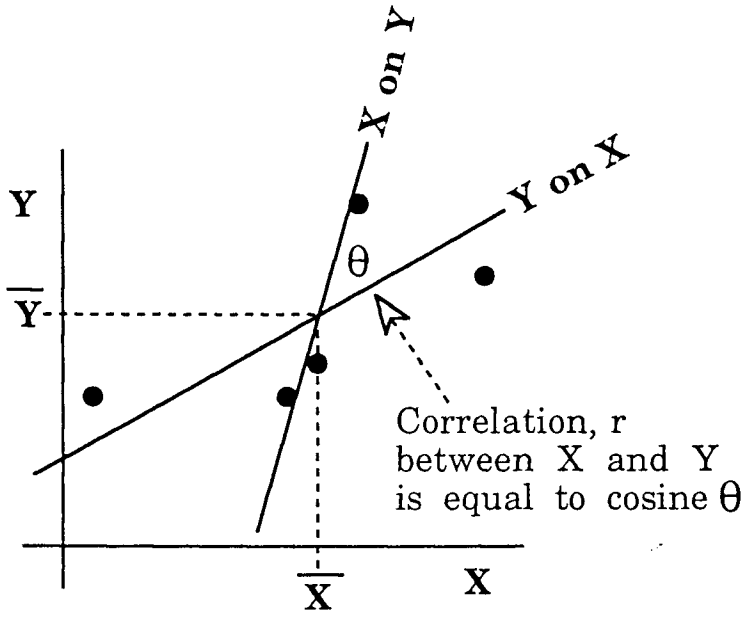
At low correlation coefficients, the divergence between the two lines is large and neither of them appears to be a "best fit" as seen by the human eye. A visually pleasing line would generally be chosen at a position that approximately bisects the two lines, because the human tends to minimize the spread perpendicular to the line, rather than parallel to either of the axes. This solution is matched closely by the reduced major axis (RMA) line (see RELATIONSHIP BETWEEN REGRESSION AND RMA LINES).

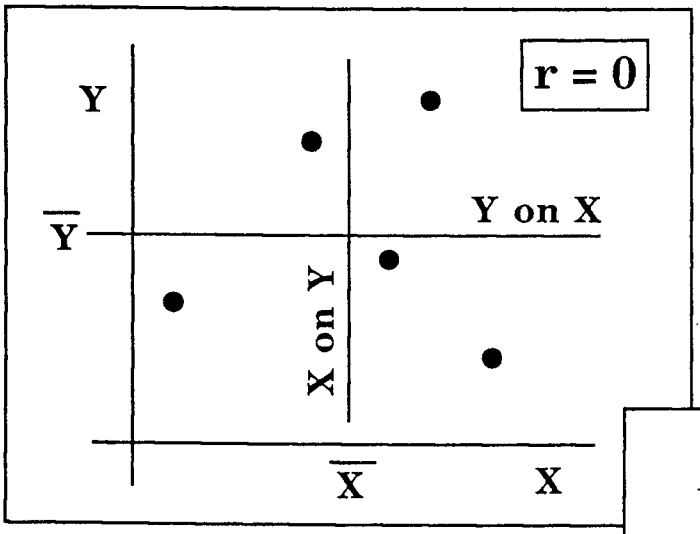
In common with both of the regression lines, the RMA passes through the bivariate mean: $\overline{X},\overline{Y}$. The slope of the line is the ratio of the standard deviations of X and Y: $slope = \frac{s_y}{s_x}$. Because the standard deviations are always positive, and the slope can be either positive or negative, the sign of the slope is given by the sign of the Pearson correlation coefficient (see REDUCED MAJOR AXIS).

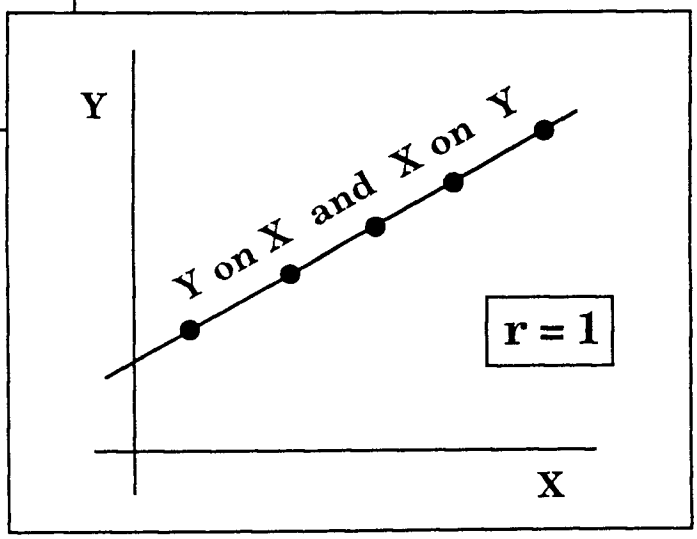
The line generally "looks good" and has been widely used as an alternative to either of the regression lines. Basically, the model attributes equal error magnitude to the variables, x and y. This may, or may not, be true. A disquieting feature of the RMA is that, unlike the regressions, its computation does not consider the covariation between x and y. The same RMA could be computed for two sets of data, both with the same means and variances, but with radically different correlations (provided that they had the same sign).

Clearly, there are often some tricky decisions to be made and these are probably best discussed in the context of a real, practical, and potentially very economically sensitive example. The following text explores regression analysis applied to the calibration of acoustic velocity logs to core porosity measurements.

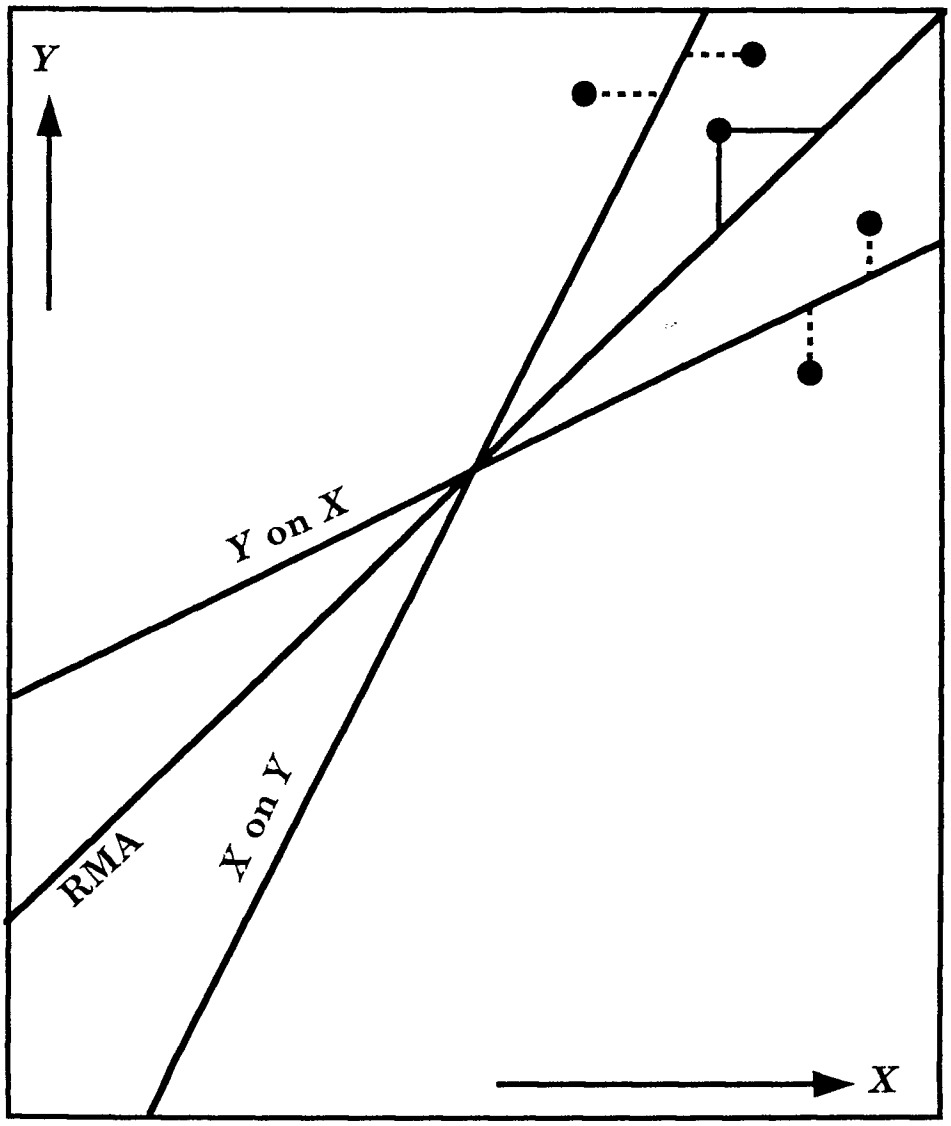
COMPARISONS OF REGRESSION OF Y - ON - X AND X - ON - Y



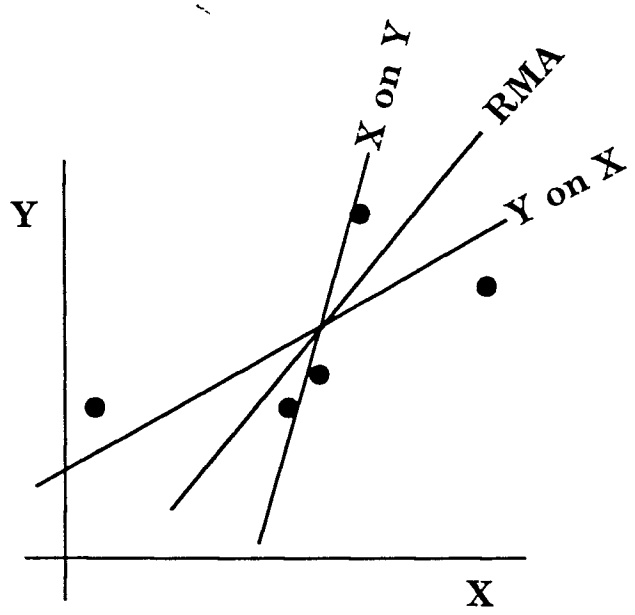


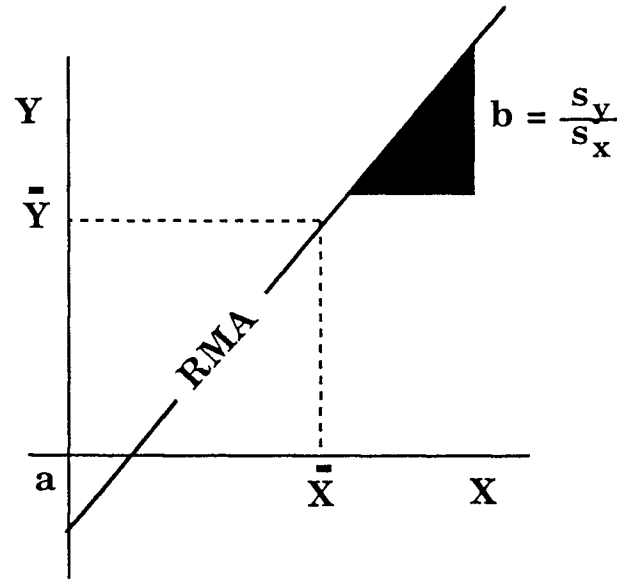


RELATIONSHIP BETWEEN REGRESSION AND RMA LINES



THE REDUCED MAJOR AXIS





Slope = Y standard deviation divided by X standard deviation (S_Y / S_X)

The sign of the slope is the same as that of the correlation coefficient.

The intercept can then be calculated, because the RMA line passes through the mean:

$$\bar{Y} = a + b\bar{X}$$

Heseldin (1968) recommended the use of the error ratio in least squares fitting of data from log analysis and demonstrated the improvement in performance when compared with standard regression or other line-fit procedures.

How can one determine the error variances or even estimate λ in practice? Collins and Pilles (1980) pointed out that random error of logging data is apparent when contrasting repeat runs of properly calibrated instruments with main runs, provided that there is no bias in the measurement. If there is a distinctive bias, then this is the concern of standard quality control procedures, where differences in the main and repeat sections reveal systematic effects that can be attributed to either tool problems, depth discrepancies, or poor hole conditions. Good examples of the recognition of logs with these systematic errors were described by Farnan and McHattie (1984), based on their extensive experience in the digital comparisons of repeat and main runs. Logs of acceptable quality have errors with a relatively small unbiased scatter that is a function of the physics of the tool, its response characteristics, and the borehole environment. In particular, the nuclear tools are subject to statistical counting error, because they record stochastic atomic processes of radioactive decay and particle generation. By contrast, electrical measurements are deterministic, but are still subject to error, determined by the precision of the instrument under borehole conditions.

In many instances, data will not be readily available to compute the error variances directly. However, they can be estimated approximately, if some concrete notion of precision can be associated with each variable. So, for example, if a known resolution, U, of a measurement device can be considered as equivalent to a 95% statistical confidence limit for the observed value, then the error variance is:

$$E^2 = \left(\frac{U}{1.96}\right)^2 = 0.26U^2 \quad \text{(Mark and Church, 1977)}$$

because 95% of the normal distribution is contained within 1.96 standard deviations of the mean. When the resolution of a variable measurement is a matter of opinion based on experience, then the numbers in this formula are themselves overly precise! However, the form of the equation gives a useful rule-of-thumb guide to the effect that the error variance is about a quarter of the squared resolution.

When no data analysis or prior knowledge can be brought to bear on the problem of error variance, then the error variance ratio, λ, is often estimated following one of two assumptions. The first assumption considers that the best estimate of the error variance ratio of two variables is given by the ratio of their total variances i.e $\lambda = \frac{s_X^2}{s_Y^2}$ This choice was advocated by Dent (1937) as the

maximum likelihood estimate of λ in the absence of any other information. The method minimizes the areas between the points and the best-fit line, which is known as the reduced major axis (RMA). In common with the other best-fit lines, the reduced major axis passes through the bivariate mean and its slope is simply the ratio of the standard deviations of the two variables, with the appropriate sign given by that of the correlation coefficient.

The alternative second assumption states that the error variance ratio is unity, i.e. $\lambda = 1$. This stipulation implies that the two variables have equal errors if the measurements are made in the same units. The best-fit line that is generated by this assumption is the principal axis, which minimizes the squared deviations, measured perpendicular to the line. It corresponds to the principal eigenvector of the variance-covariance matrix. Unlike the other best-fit lines, the solution is sensitive to the units of measurement. Consequently, the principal axis is usually calculated for standardized data and then transformed to the original units.

In summary, the choice of best-fit line is first determined by the purpose of the procedure. If the intent is only to make predictions of one variable on the basis of measurements of another, then regression is the preferred choice. The variable to be predicted is the dependent variable, the predictor is the independent variable. Alternatively, when a functional analysis is the goal, and where the controlling parameters have both meaning and utility, the best-fit line should incorporate estimates of the random errors associated with each variable. Wherever possible the error variances should be computed from replicate samples, which in the case of wireline logs are provided by the consideration of both main and repeat runs. At the other extreme, the error variance ratio can be assumed to be linked with the total variance in the computation of either the reduced major axis or the principal axis.

Yet another option is available in functional analysis, when it is realized that the selection of the most appropriate line-fit should provide the most reasonable error variance ratio and simultaneously, the equation intercepts and slopes that match the rock properties and physical constraints of the functional relationship.

Although this information was not available for the "Equity Sandstone",the error variances could have been estimated as a contributory part of the line-fit analysis. The error variance of the transit times would be estimated by analysis of the deviations of the main sonic log from its repeat section, in a similar procedure to that applied to the density log example described earlier. The error variance of core samples is comprised of two sources of variability. The first is controlled by the resolution of the laboratory method of porosity measurement, which can be deduced from repeated analysis of the same core samples. This procedure is a fundamental quality check and is widely practised by laboratories on standard core samples to gain information on relative precision, and to check for bias when comparing with alternative methods or different laboratories. The

results of this type of work are now reported more widely, such as the statistical summary of the data quality assurance tests at Amoco described by Thomas and Pugh (1989). However, the integration of such data as a part of standard log analysis is still a rare event. The second source of core variability is caused by the fact that measurements are most commonly made on small plugs sampled at intervals of one foot. These are only estimates of the porosities represented in whole-core measurements. The smaller volume causes plug measurements to have higher variances than those of the larger whole-core samples.

In the absence of specific information on the error variances of the two variables, functional analysis proceeds by evaluating the consequences of alternatives in search of an optimum line-fit. The criteria to be met are that the joint estimates of the functional parameters and the error variance ratio, λ should be judged the most reasonable combination. The range of possibilities are bound by the regression line at each extreme, where the total error is attributed to one or other of the variables. However, it should be remembered that these extremes are only estimates of the true regression lines, because they are based on a sample size of 44 observations. All the possible best-fit lines pass through the bivariate mean, so that the trace of possible parameter solutions is a straight line as shown on the crossplot of matrix and fluid transit time of Figure 11. The reduced major axis (RMA) is close to the idealized transit times of quartz and fresh water. If this line is the best choice, then the error variance ratio would be 0.52. Converting transit times to porosity equivalences, the number would suggest that the sonic log and core data estimate the porosity to about the same accuracy. This conclusion is credible when it is remembered that the core measurement is based on a small plug and is only an estimate of the whole rock sample. This would probably cause the standard deviation of both estimates to be approximately the same at about one porosity unit.

REFERENCES

Collins, H. N., and D. Pilles, 1979, Some uses of functional analysis in petrophysics: Canadian Well Logging Society 7th Annual Symposium, Paper E, 17 p.

Dent, B. M., 1935, On observations of points connected by a linear relation: Proceedings of the Physical Society of London, v. 47, pt. 1, p. 92-108.

Farnan, R. A., and C. M. McHattie, 1984, Use of digital overlays and crossplots for log quality evaluation: The Log Analyst, v. 25, no. 1, p. 3-10.

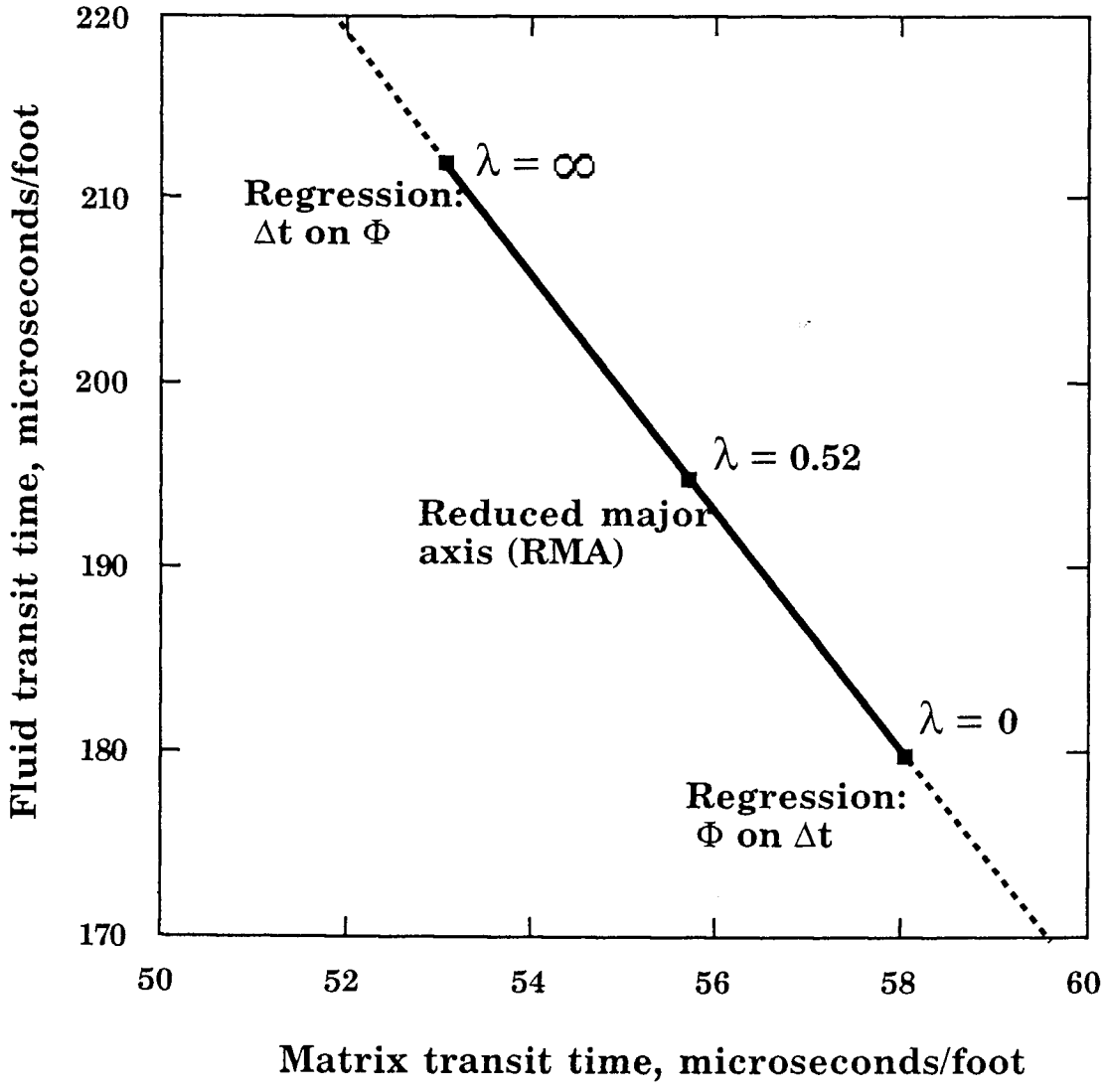
Heseldin, G. M., 1968, The use of error ratio in least square fitting of data: The Log Analyst, v. 9, no. 3, p. 22-25.

Mark, D. M., and M. Church, 1977, On the misuse of regression in Earth science: Mathematical Geology, v. 9, no. 1, p. 63-75.

Thomas, D. C., and V. J. Pugh, 1989, A statistical analysis of the accuracy and reproducibility of standard core analysis: The Log Analyst, v. 30, no. 2, p. 71-77.

FUNCTIONAL ANALYSIS: CROSSPLOT OF THE MATRIX AND FLUID TRANSIT TIMES FOR THREE BEST-FIT LINES

(λ is the error variance ratio)



MODELS AND REALITY

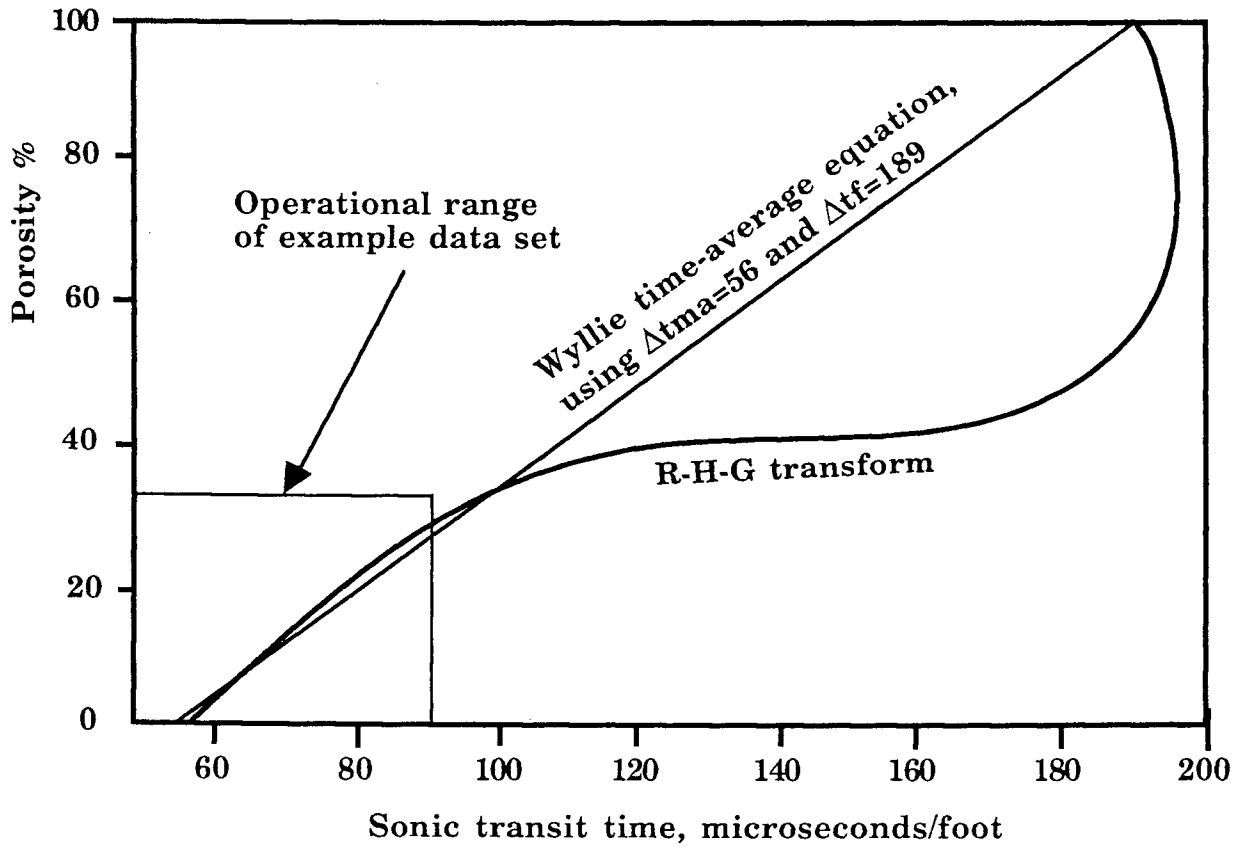
When evaluating the results of functional analysis, clear distinctions must be made between useful descriptive models and functional relations that are actual mathematical descriptions of processes. In this present example, the Wyllie equation is a descriptive functional relationship that should only be honored to the extent that it models reality. The shortcomings of the time-average relation were understood at the outset by Wyllie et al (1956), and modifications have been proposed over the years, of which the most widely adopted is that of the Raymer-Hunt-Gardner transform. From a study of many sandstones, Raymer, Hunt and Gardner (1980) established a generalized transit time - porosity relationship (see THE RAYMER-HUNT-GARDNER...). Since the curve is a closer representation of functional reality than the equation used up to now, how does this affect our conclusions? Examination of the figure shows that for the data range of the example, the curve is closely approximated by the time-average equation, with expectations of matrix and fluid transit time that would be close to their physical values. However, if the samples had been drawn from a higher porosity range, then a linear trend would have been tangential to the curve, with an expected apparent matrix transit time artificially lower than its real value. Consequently, the expectations of credible parameters would need to be modified appropriately. These considerations do not invalidate the approach of functional analysis, but instead remind the analyst that reality takes precedence over models when they reach their limitations. It should also be noted that this same warning applies to the Archie equation and several other fundamental log analysis relationships.

REFERENCES

Raymer, L. L., E. R. Hunt, and J. S. Gardner, 1980, An improved sonic transit time-to-porosity transform: Transactions of the SPWLA 21st Annual Logging Symposium, Paper P, 12 p.

Wyllie, M.R.J., Gregory, A.R., and Gardner, L.W., 1956, Elastic wave velocities in heterogeneous and porous media: Gephysics, v.21, no.1, p. 41-70.

THE RAYMER-HUNT-GARDNER (R-H-G) SONIC TRANSIT TIME-TO-POROSITY TRANSFORM CONTRASTED WITH THE WYLLIE TIME-AVERAGE EQUATION



ESTIMATION OF POROSITY FROM SONIC LOG TRANSIT TIMES IN THE "EQUITY SANDSTONE" (From Doveton, 1994)

In this example, we examine the problem of transforming transit times from a sonic log to a porosity equivalent, using core measurements of porosity. The consequences of the choice of one or other of alternative line-fits are by no means of purely academic interest. It is now common practice for estimations of porosity to be tied to core - log calibrations in unitized fields. By this means, porosities can be calculated in uncored wells and used for estimation of volumetrics on a field-wide basis. Even minor differences in line slope can cause significant changes in the allocation of reserves between the participating operators. This was widely appreciated at the equity hearings of the 80's when there was considerable debate as to the relative merits of alternative statistical line-fit strategies.

The data consist of 44 measurements of sonic log transit time (Δt) of a sandstone reservoir, matched with core porosities (ϕ)at equivalent depths (see CORE POROSITIES...). The core porosities were previously smoothed by a moving average filter, because they were sampled at one foot increments and the measurement span of the sonic log was of length two feet. This initial remedial step ensures an approximately common vertical resolution between the two measurement types. Failure to do this results in data incompatibility, which causes both distinctive error and bias.

REFERENCE

Doveton, J.H., 1994, Geologic Log Analysis Using Computer Methods: AAPG Computer Applications in Geology, No. 2, 169 pp.

CORE POROSITIES (PERCENT) AND SONIC LOG TRANSIT TIMES (MICROSECONDS PER FOOT) FROM THE EQUITY SANDSTONE

Φ	Δt	Φ	Δt
6.8	63.8	13.4	75.1
9.3	66.6	13.4	74.4
8.5	68.1	13.4	72.1
10.9	68.3	13.5	69.9
10.3	69.9	14.3	72.8
10.4	70.5	15.1	72.9
10.2	71.6	15.1	74.5
10.1	72.2	15.1	75.1
10.1	72.7	15.3	77.9
8.6	72.2	15.2	78.5
11.0	72.7	15.2	80.0
11.0	72.3	15.2	81.0
10.9	71.6	16.9	83.8
11.0	71.1	17.6	81.1
11.1	70.0	17.6	79.5
11.7	70.9	16.9	76.0
11.8	71.6	17.8	77.3
11.8	73.3	18.6	76.7
11.8	73.9	18.5	79.1
11.8	74.8	20.1	82.3
12.6	75.6	20.3	83.5
12.7	73.8	19.6	84.5

ALTERNATIVE BEST-FIT LINES OF POROSITY AGAINST SONIC LOG TRANSIT TIMES IN THE "EQUITY SANDSTONE" (From Doveton, 1994)

Initially, the problem can be seen to be one of simple prediction : given a transit time from a sonic log, what is the porosity of the zone, if it was cored and analyzed? A linear relationship between porosity and transit time is commonly assumed to be a usable approximation (Wyllie et al, 1956). The prediction equation is then:

$$\hat{\phi} = a + b\Delta t$$

which is a regression of porosity on transit time. The result corresponds to the most shallowly sloping line on the crossplot of porosities and transit times (see BEST-FIT LINES...). The regression equation is:

$$\hat{\phi} = -33.3 + 0.63\Delta t$$

and has a coefficient of determination of 0.76, meaning that the linear prediction accounts for 76% of the total variability, with the remaining 24% left in the residual squared deviations about the line. The coefficient of determination is equal to the square of the correlation between core porosity and transit time, which is 0.87.

This regression model of porosity on transit time ascribes all the error to the core porosity and none to the transit time. The consequences can be seen on RESIDUAL DIFFERENCES... where the errors in predicted porosity are random when graphed against transit time, but show a tendency for underprediction at the high end and overprediction at the low end when plotted against measured porosity. This effect simply shows that any prediction of porosity is the best on average for any given value of transit time. However, as pointed out by Collins (1984) the choice of this line would be resisted in unit operating negotiations by an owner whose property had porosities that tended to be higher than the average.

The alternative regression of transit time on porosity results in the steepest line fit (see BEST-FIT LINES...) and allocates all the error to the transit time with none to core porosity. The descriptive equation is:

$$\hat{\Delta t} = 58.1 + 1.2\phi$$

This line-fit solution would be welcomed by a property owner with higher than average porosity for the opposite reasons attached to the other regression line : now the result would appear to enhance higher porosities, while further

downgrading lower porosities. Traditional regression texts would reject this alternative out of hand, since they view the situation as one of predicting the best estimate of porosity on average, based on a given value of transit time. However, others would argue that this is a calibration problem in which the core porosities must be honored as the calibration standard, and so effectively considered as free of error.

Finally, the prediction line that is often used as a compromise between the two regression extremes is the reduced major axis (RMA). The line passes through the bivariate mean and its slope is determined by the ratio of the standard deviations of the core porosity and transit time. The equation of the RMA line is then:

$$\hat{\phi} = -40.0 + 0.72\Delta t$$

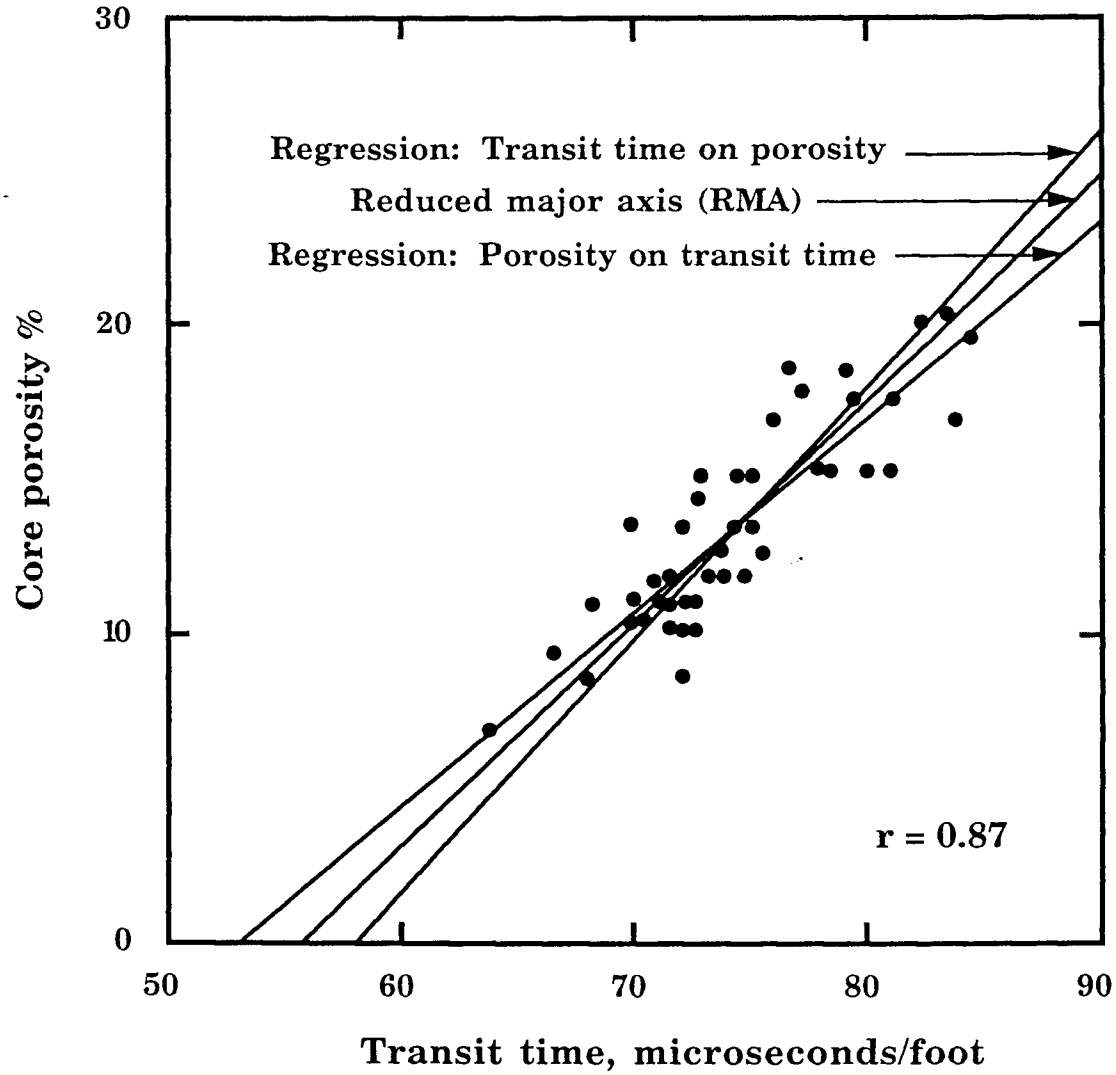
The selection of the RMA is sometimes based on intuitive appeal, since it typically has the visual appearance of best fit. The reason for this is that a best-fit line drawn by eye will usually minimize scatter about the line in a direction normal to the line. This is the criterion for the principal axis, but is also closely approximated by the reduced major axis. The comparative simplicity of the equation parameters and its failure to include cross product terms between the two variables puts a strain on its credibility. However, if the measurement error variance ratio is closely approximated by the total variance ratio then the RMA will be the optimal solution.

REFERENCEs

Collins, H.N., 1984, Regression analysis - some loose ends: Canadian Well Logging Society Journal, v.13, no. 1, p.61-64.

Wyllie, M.R.J., Gregory, A.R., and Gardner, L.W., 1956, Elastic wave velocities in heterogeneous and porous media: Gephysics, v.21, no.1, p. 41-70.

BEST-FIT LINES TO THE EQUITY SANDSTONE CORE POROSITY - SONIC LOG TRANSIT TIME DATA



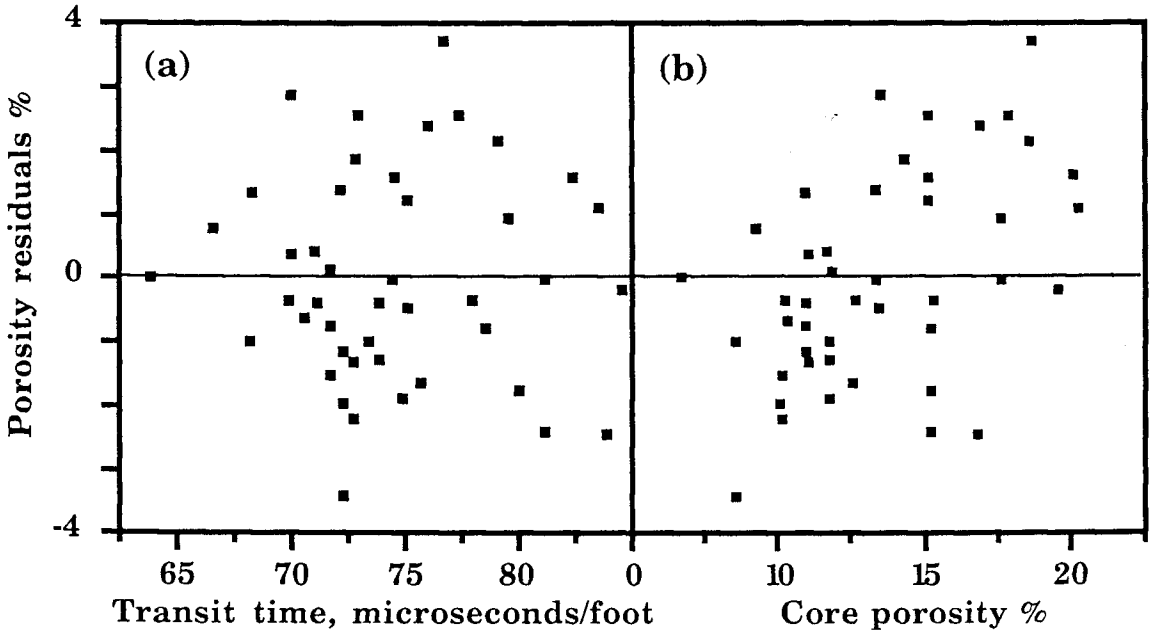
Equations of the alternative best-fit lines:

Φ on Δt: $\hat{\Phi} = -33.3 + 0.63\ \Delta t$

Δt on Φ: $\hat{\Delta t} = 58.1 + 1.2\Phi$

RMA: $\hat{\Phi} = -40.0 + 0.72\ \Delta t$

RESIDUAL DIFFERENCES BETWEEN CORE POROSITIES AND PREDICTIONS BASED ON THE REGRESSION OF POROSITY ON THE TRANSIT TIME VERSUS (a) TRANSIT TIME AND (b) CORE POROSITY

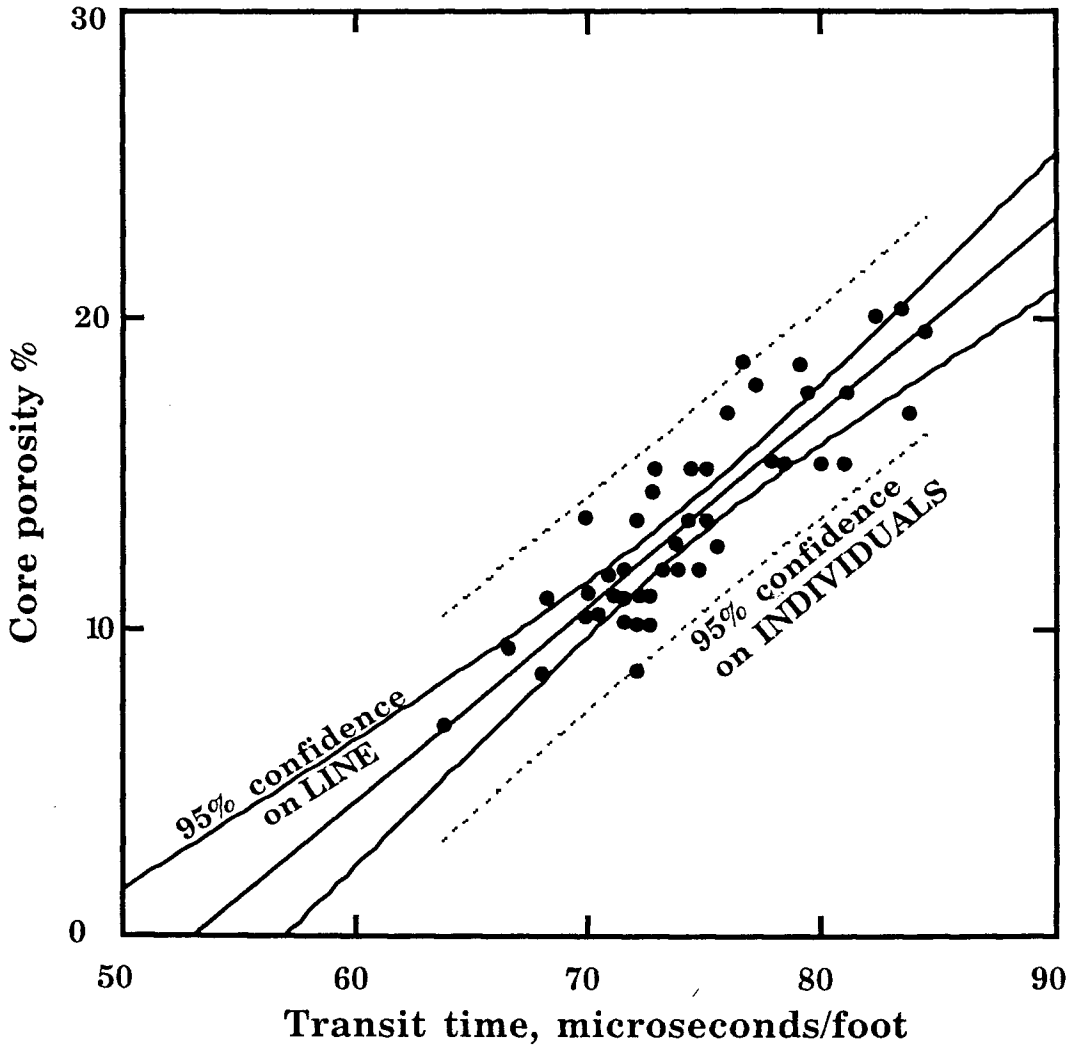


CONFIDENCE LIMITS

A regression line is estimating the AVERAGE value of Y given any particular value of X. The residuals (or deviations) are modeled as normally distributed error about the line parallel to the Y axis. Confidence limits for INDIVIDUAL observations may be computed as belts on either side of the line (see dashed lines in 95% CONFIDENCE BELTS...). If a 95% level is chosen, then we would expect 95% of the data points to plot within these bounds. The Equity Sandstone data consists of 44 points and suggests a general expectation that about two points should lie outside the bounds if the residuals are normally distributed. The plotting of these confidence belts is a good means to highlight "outliers" as possibly freak observations that need further evaluation (and possible elimination) in the quality control of data sets.

Confidence belts may also be computed on the LINE itself, because the line is calculated as a single estimate of the true population parameter line. By computing (say) 95% confidence belts on the trend, we can establish a zone within which we can be 95% confident that the true parameter line would occur, if we had infinite observations. Visually, the zone gives an idea of the relative "play" on the estimated line (see solid curves in 95% CONFIDENCE BELTS...).

95% CONFIDENCE BELTS AND STATISTICS OF REGRESSION OF POROSITY ON SONIC TRANSIT TIME IN THE EQUITY SANDSTONE



Linear Fit

Summary of Fit

Rsquare	0.764317
Root Mean Square Error	1.663504
Mean of Response	13.46591
Observations (or Sum Wgts)	44

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	376.91450	376.914	136.2056
Error	42	116.22436	2.767	Prob>F
C Total	43	493.13886		0.0000

FUNCTIONAL ANALYSIS

Because the correlation between logging variables can often be moderate to low, evaluation of the coefficients of the alternative regression equations can be problematical if the equations have a functional petrophysical meaning. When the relationship between the two variables is subject to a proven (or at least, accepted) physical model or known natural constraints, the goal becomes functional analysis, rather than simple prediction. The two regression lines are then seen to be the two limiting extremes, where all the error is attributed to either one or other of the two variables. The real functional line should lie somewhere in between, with its slope controlled by the relative amount of error assigned to each variable. The error in question is due to random measurement effects that result from both the tool characteristics and the fluctuations in the borehole environment. The issue in question is the precision of the measurement rather than its accuracy.

The error variance of a variable X gives measurement precision and can be determined by repeating the measurement for n replicates of the same observation, when:

$$E_X^2 = \frac{\Sigma(X_i - \overline{X})^2}{n}$$

If the error is independent, then the error variance can be determined for each variable separately. The ratio between error variances:

$$\lambda = \frac{E_Y^2}{E_X^2}$$

can be used to estimate the true functional line that takes into account the relative amount of measurement error associated with both variables. When $\lambda=0$, then the Y values are known without error and the appropriate solution is an X-on-Y regression. At the other extreme, when λ is infinite, all the error is linked with the Y variable and the choice must be a regression of Y-on-X. In all intermediate cases, the line will be located between and its slope can be calculated from the slope of the Y-on-X regression by:

$$b_f = \frac{\left(\frac{b^2}{r^2} - \lambda\right) + \sqrt{\left(\frac{b^2}{r^2}\right) + 4\lambda b^2}}{2b}$$

COMPATIBILITY OF VERTICAL RESOLUTION

Comparisons of petrophysical variables must be made in terms of a common vertical resolution. The necessity for this rule has been discussed widely in the log analysis literature so that, for example, Runge and Powell (1967) stated that: " Different logging devices and sampling techniques have different spans and when a comparison is desired, the differences in span lead to an incompatibility between these modes of measurement". Ideally, a deconvolution of the coarser resolving measurement to a finer scale would be desirable, but this is generally not practical. Problems such as the non-linear response of the induction tool and the stochastic nature of nuclear measurements make effective deconvolution very difficult (Looyestijn, 1982). In practice, measurements are smoothed to an equivalent common scale with the variable with the coarsest vertical resolution. In the current example, the core measurements of porosity were smoothed by a running-average filter to give an approximate common resolution with the two foot span of the sonic log.

There are consequences that follow from the failure to correct the incompatibility of measurement scale by appropriate smoothing. These can be better understood by consideration of the relationship between porosities of plugs and whole-core samples. In experiments with the early density tool, Baker (1957) contrasted porosities measured from one-inch diameter plugs with porosities measured from their whole core samples. A crossplot of the results are shown on CORE PLUGS v. WHOLE CORE ... and show that the error in predicting the porosity of any one foot interval on the basis of a plug measurement has a highly distinctive bias. At higher porosities, the plug will tend to overestimate the average porosity, while at lower values the plug will tend to be an underestimate. The relationship is inevitable, because the porosity of the whole core represents an average of all the potential plugs it contains. Consequently, although a set of plugs and whole core should have the same mean value, the variability of the whole core will be less than that of the smaller plugs. The decreased variance on the whole-core porosity axis, compared with the variance on the plug porosity axis, gives the appearance of rotating the data cloud from a simple diagonal trend. The mechanism for this effect is simply the aggregation process of measurements from larger volume samples, in which the extremes in the smaller samples are averaged out. Although these arguments have been developed from the perspective of core measurements, they apply equally to logging data.

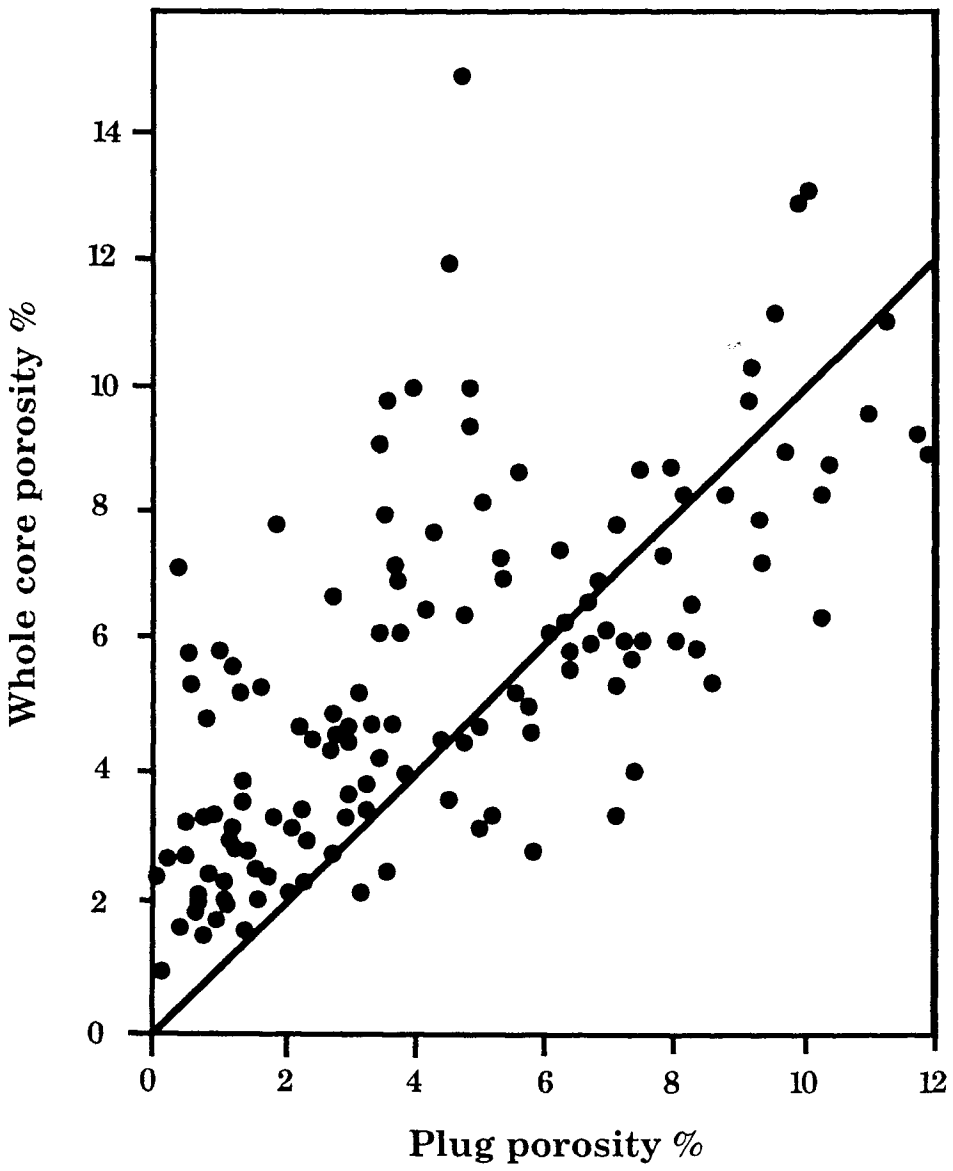
REFERENCES

Baker, P. E., 1957, Density logging with gamma rays: Petroleum Transactions of the AIME, v. 210, no. 3, p. 289-294.

Looyestijn, W. J., 1982, Deconvolution of petrophysical logs: Applications and limitations: Transactions of the SPWLA 23rd Annual Logging Symposium, Paper W, 20 p.

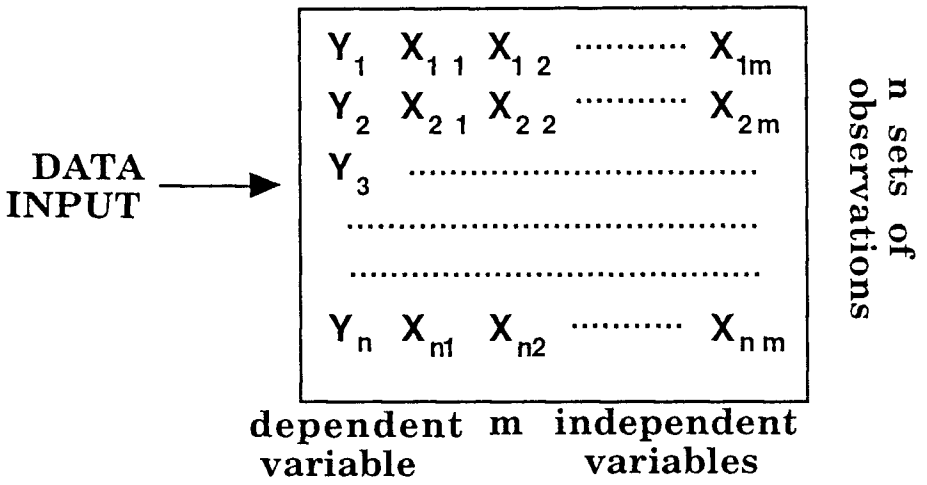
Runge, R. J., and N. J. Powell, 1967, The effect of sampling on sonic log span adjustment: Transactions of the SPWLA 8th Annual Logging Symposium, Paper D, 14 p.

CORE PLUGS V. WHOLE CORE COMPARISON OF POROSITIES. Adapted from Baker (1957)



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 :



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 :

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \hat{Y}_m \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots\ldots & X_{1m} \\ 1 & X_{21} & X_{22} & \ldots\ldots & X_{2m} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 1 & X_{n1} & X_{n2} & \ldots\ldots & X_{nm} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ a_m \end{bmatrix}$$

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

Now, the solution is found by minimizing the sum of squares deviations between Y and $\hat{Y}_i$, 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: AN ESTIMATION OF PERMEABILITY FROM LOGS EXAMPLE

The most simple quantitative methods to predict permeability from logs have been keyed to empirical equations of the type :

$$K = A\Phi^B$$

where A and B are constants determined from core measurements, and applied to log measurements of porosity (Φ) to generate predictions of permeability (K). When applied to special cases of homogeneous sandstones, the results may be adequate, but prediction errors are often large in typical sandstones, and the errors in predicted permeability commonly range across orders of magnitude when applied to carbonates. The reason for this is that permeability is not exclusively determined by pore volume, but is also controlled by internal surface area, pore network tortuosity, pore throat geometry and other variables.

Wyllie and Rose (1950) developed an equation which linked permeability with both porosity and irreducible water saturation (S_{wi}), based on laboratory measurements of core:

$$K = \frac{A\Phi^B}{S_{wi}^C}$$

The rationale of this equation can be understood when it is compared with the classic Kozeny-Carman equation :

$$K = \frac{A\Phi^3}{(1-\Phi)^2 S^2}$$

which incorporates the specific surface area, S. The specific surface area is the ratio of surface area to volume of framework solid and is difficult to measure directly by conventional methods. However, the specific surface area is inextricably linked with pore size, which in turn controls irreducible water saturation. The irreducible water saturation term in the Wyllie-Rose equation therefore functions as a powerful surrogate variable for specific surface area. When applied by Timur (1968) to sandstones, the results showed a considerable improvement in permeability estimation over those based on porosity values alone.

These ideas can be extended to carbonates in models which incorporate concepts drawn both from depositional facies and diagenetic processes. Several log measures should be useful, particularly since diagenesis is often fabric-selective and frequently linked with changes in mineral composition. The following

example uses a data set of core permeabilities and logging measurements from the Lower Permian Chase Group of the giant Hugoton gasfield in southwest Kansas (Doveton, 1994). The raw measurements of permeability were smoothed with a 5-point (two and a half foot) binomial filter to give approximately equivalent vertical resolution with the wireline logging measurements. The data set consists of zoned readings of permeability, porosity computed from a density-neutron log combination, uranium and potassium measures from a spectral gamma-ray log.

The regression line of permeability on porosity for the Chase Group data is shown in SIMPLE LINEAR REGRESSION... The line picks up the broad trend of increasing permeability with porosity, but accounts only for 14% of the total variability. The low fit causes the slope to be markedly shallow, with an accentuation of the innate tendency to underpredict high permeabilities and overestimate low permeabilities.

Multiple regression is an extension of simple linear regression analysis that incorporates additional independent variables in the predictive equation. By this means, permeability predictions may be improved through the inclusion of log measurements which are indirectly related with pore geometry, principally the internal surface area.The form of the expanded regression model is :

$$\log K = A + B * \Phi + C * L1 + D * L2 + \ldots.$$

where *L1*, *L2* etc. are additional log measurements. The choice of useful log variables is helped through the procedure of stepwise regression, where different combinations of variables are used in an iterative process to determine the set that provide the best estimate, and where the contribution of each variable is judged to be statistically significant.

In the Chase Group example, the coefficients of determination (COEFFICIENTS OF DETERMINATION...) for the alternative regressions of core permeability on all possible combinations of log measurements of porosity, uranium and potassium. There is a systematic improvement in prediction power with the inclusion of additional variables. The regression equation that links permeability with porosity and uranium represents a plane of predictions mapped on to the two dimensions of the independent variables (TWO INDEPENDENT VARIABLES...). When potassium is included as a third independent variable, the equation describes a hyperplane of predicted permeabilities in the three dimensions of porosity, uranium and potassium (THREE INDEPENDENT VARIABLES...).

The regression coefficients associated with the independent variables show a consistent pattern of increasing permeability with increasing porosity, but decreasing permeability with greater concentrations of uranium and potassium. Both of these elements are statistically significant contributors to the regression

model and so must be correlated with features of pore geometry. The potassium content appears to reflect small concentrations of illite which adversely affect permeability. The explanation for the role of uranium is more speculative, but may be linked with preferential leaching and improvement of transmissibility within the pore networks.

The porosity (ϕ), uranium (Ur) and potassium (ρ) logs were transformed into a continuous profile of permeabilities through the application of the multiple regression equation:

$$\log K = -0.07 + 0.12\phi - 0.30Ur - 1.35\rho$$

Permeability predictions outside the range of data used for the regression were discarded in order to screen out unwarranted extrapolations beyond reasonable prediction limits. The intervals eliminated by this procedure consisted of shales and shaly carbonate zones. The log is shown together with the core measurements of permeability (see PREDICTED PERMEABILITY LOG...). The match between them appears to be reasonable, although the regression accounts for only 55% of the total variability. The basic characteristic of the multiple regression as a method that tends to estimate the mean can be seen in the pattern of underestimates at higher values, overestimates of lower values of permeability.

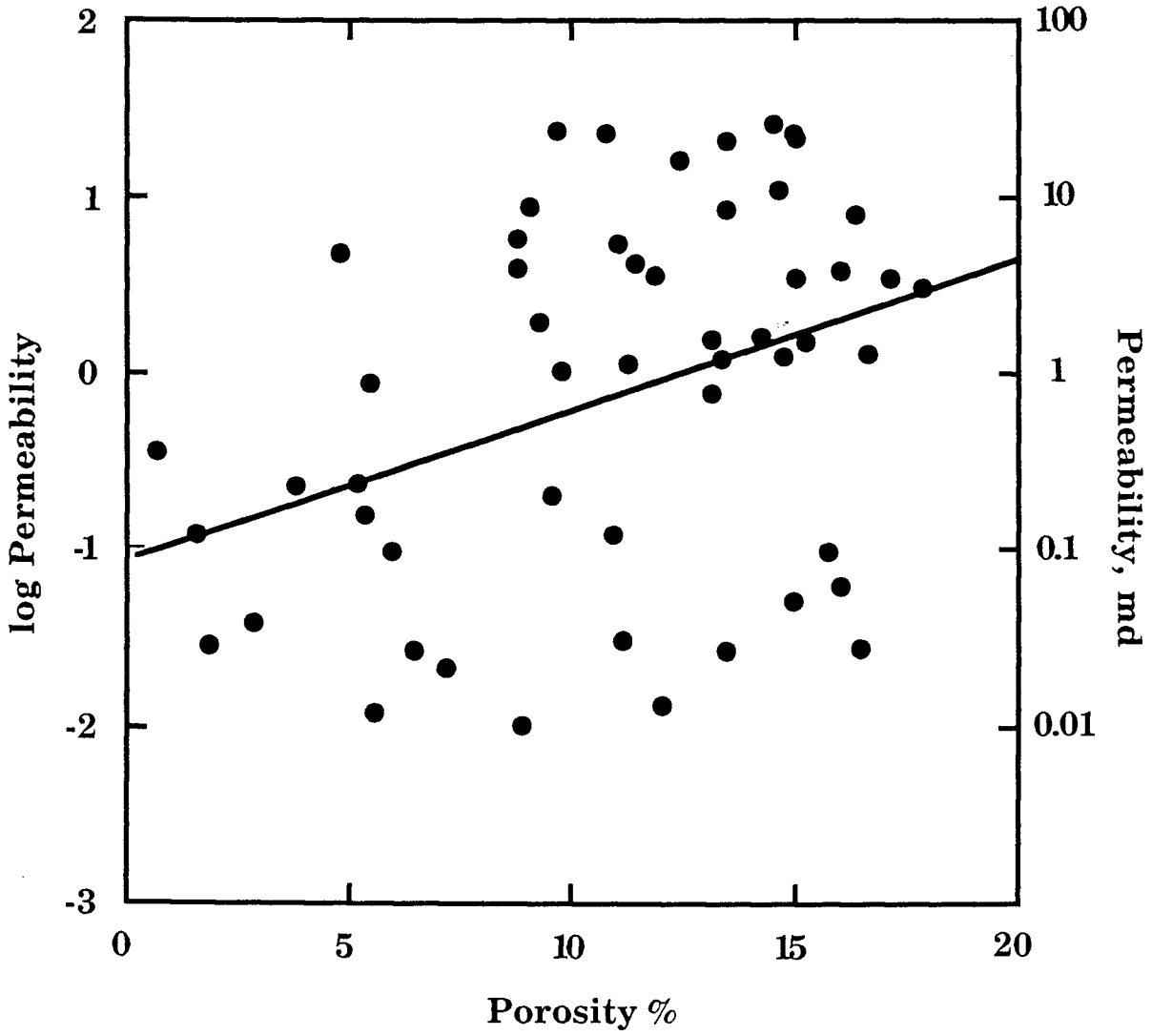
REFERENCE

Doveton, J.H., 1994, Geologic Log Analysis Using Computer Methods: AAPG Computer Applications in Geology, No.2, 169 pp.

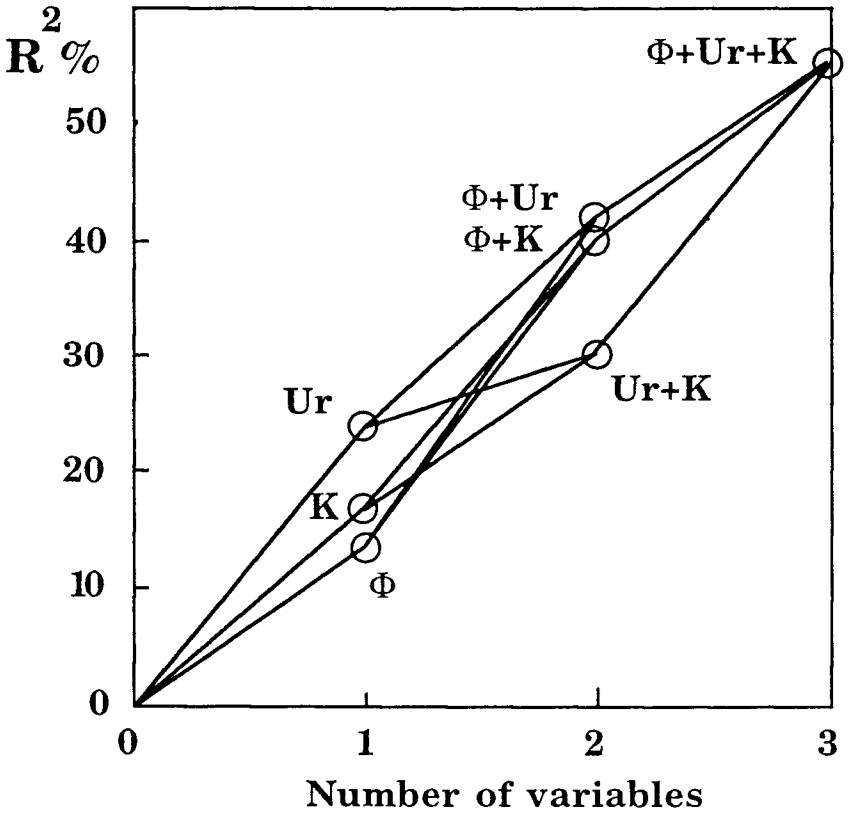
Timur, A., 1968, An investigation of permeability, porosity, and residual water saturation relationships: Transactions of the SPWLA 9th Annual Logging Symposium, Paper J, 18 p.

Wyllie, M.R.J., and Rose, W.D., 1950, Some theoretical considerations related to the quantitative evaluation of the physical characteristics of reservoir rock from electrical log data: J.Pet. Tech., p.189.

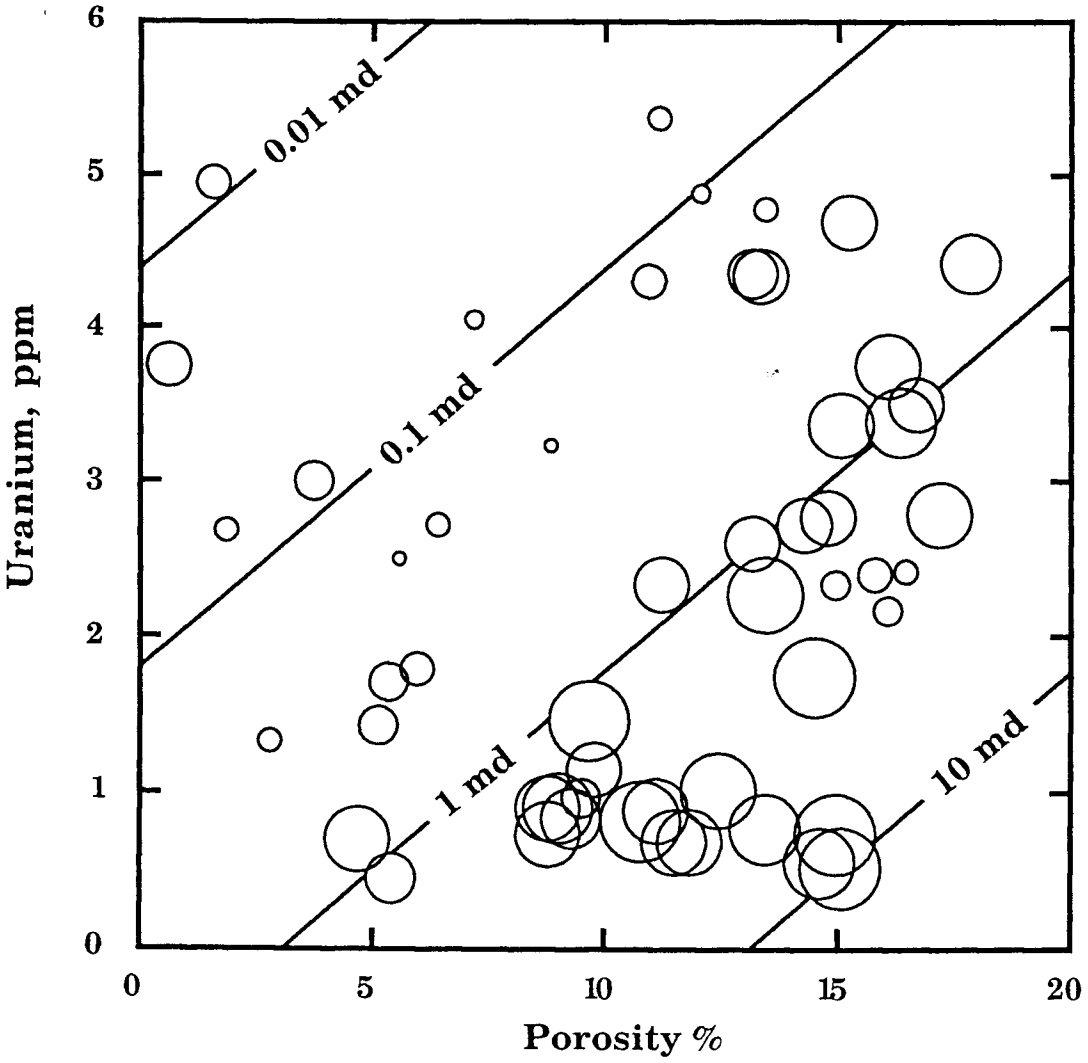
SIMPLE LINEAR REGRESSION OF PERMEABILITY ON POROSITY IN CHASE GROUP WELL



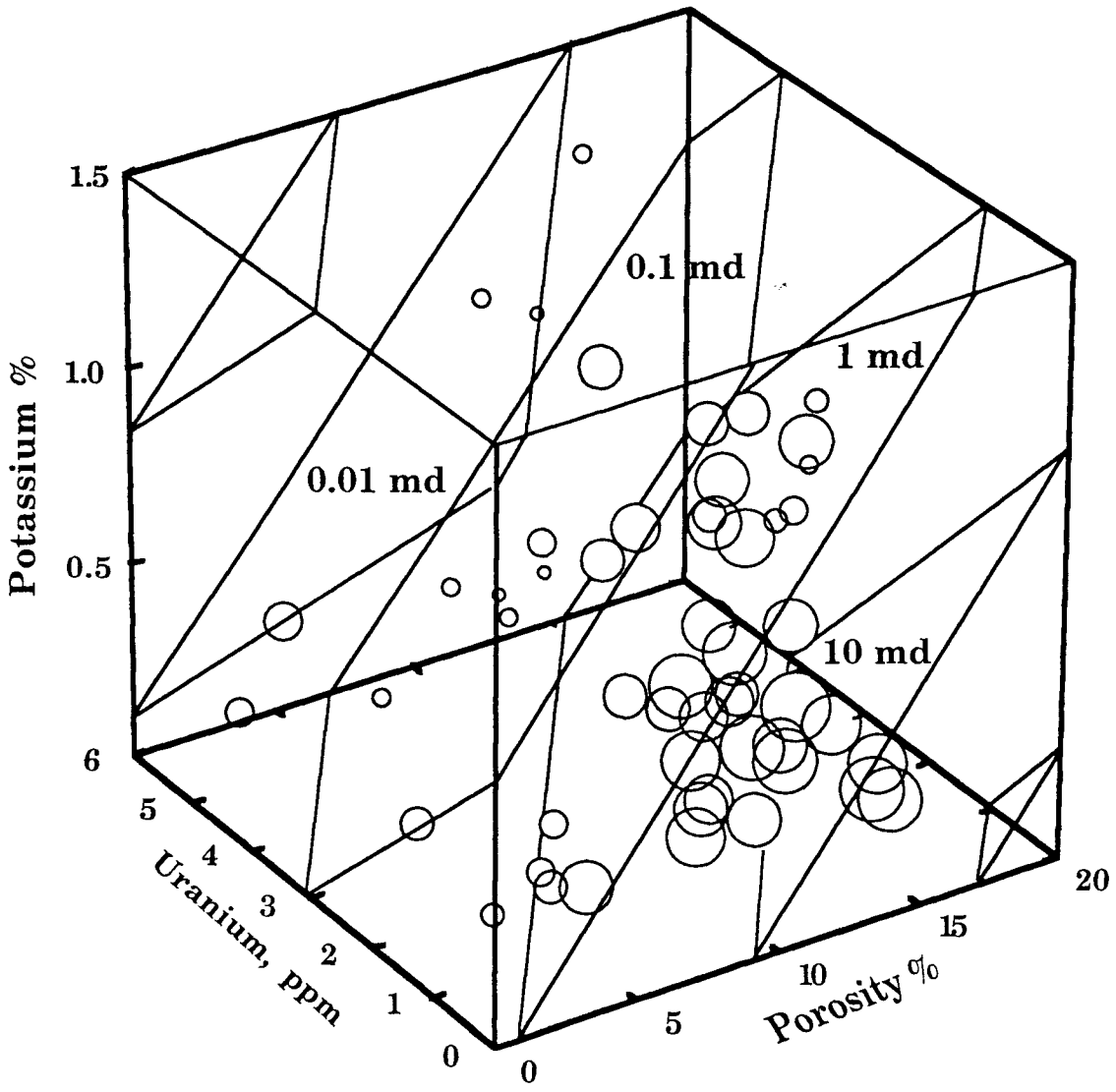
COEFFICIENTS OF DETERMINATION (R^2) FOR ALTERNATIVE MULTIPLE REGRESSIONS OF PERMEABILITY IN CHASE GROUP WELL



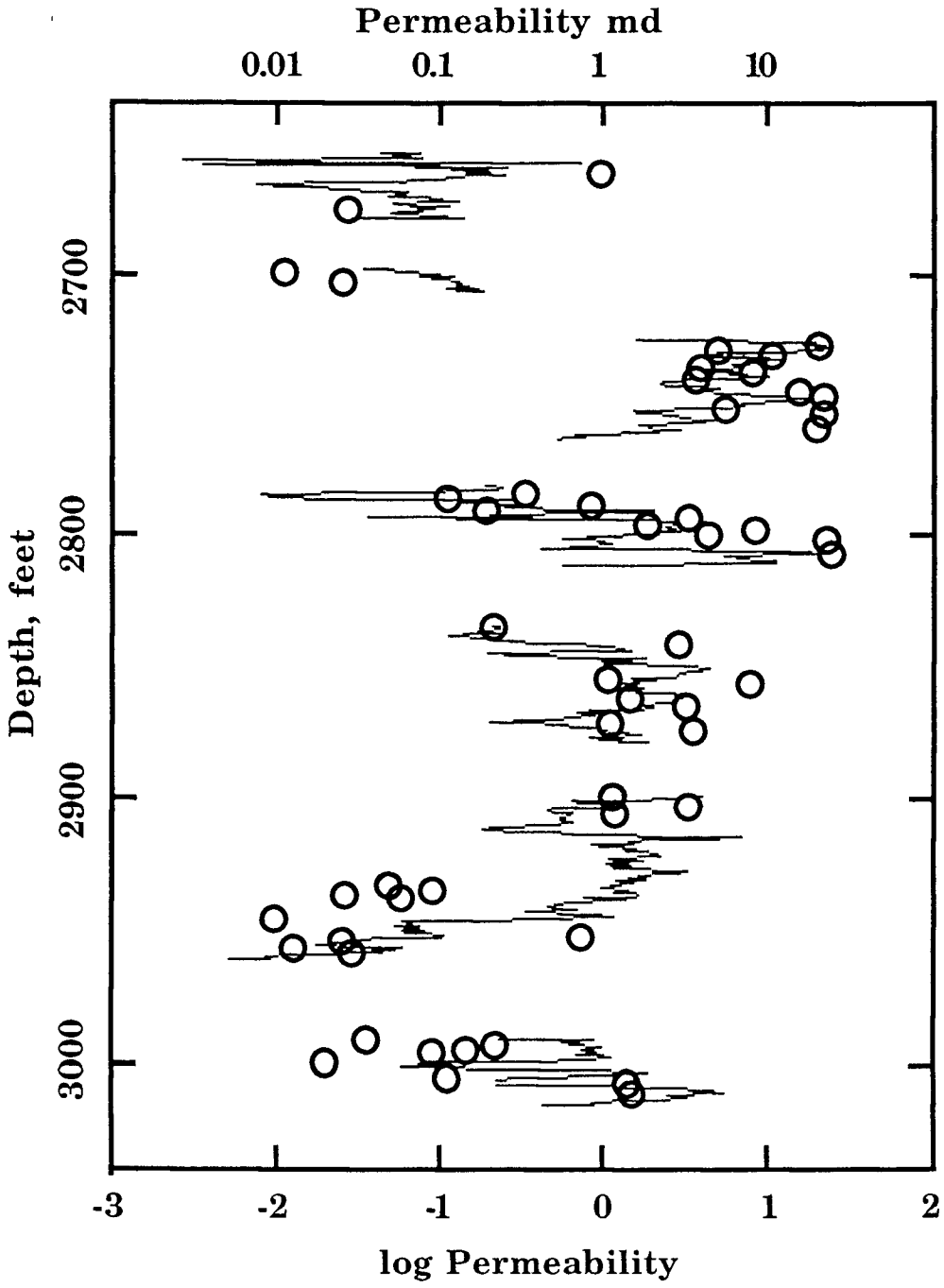
TWO INDEPENDENT VARIABLES: MULTIPLE REGRESSION OF PERMEABILITY ON POROSITY AND URANIUM IN THE CHASE GROUP WELL
Contours of regression estimates
Bubble diameters proportional to measured permeabilities



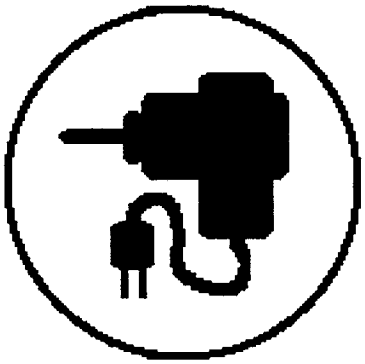
THREE INDEPENDENT VARIABLES: MULTIPLE REGRESSION OF PERMEABILITY ON POROSITY, URANIUM, AND POTASSIUM IN THE CHASE GROUP WELL
Contours of regression estimates
Bubble diameters proportional to measured permeabilities



PREDICTED PÈRMEABILITY LOG OF THE CHASE GROUP SECTION FROM MULTIPLE REGRESSION ON POROSITY, URANIUM, AND POTASSIUM
Measured core permeabilities shown by circles



MULTIVARIATE STATISTICAL POWER TOOLS

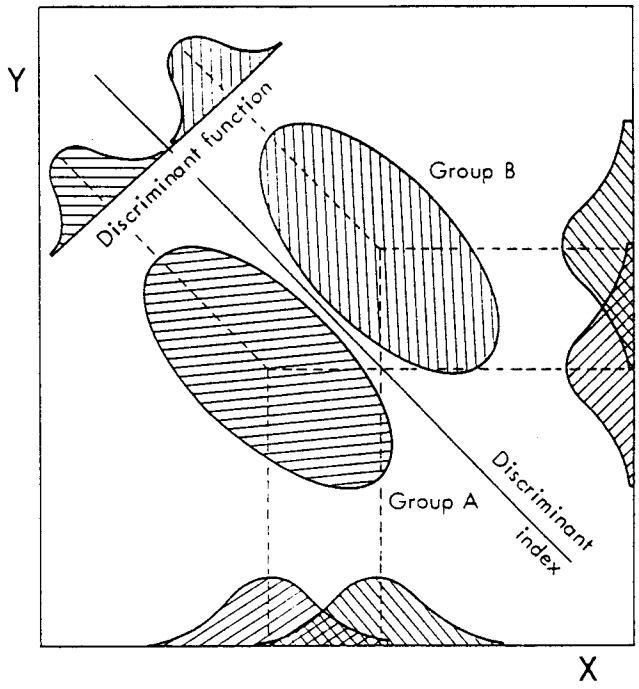


FOR

PATTERN RECOGNITION, DISCRIMINATION, AND CLASSIFICATION

LINEAR DISCRIMINANT FUNCTION ANALYSIS

A SUPERVISED statistical classification technique. A linear discriminant function is computed which best distinguishes two groups of known assignment on the basis of several observational variables. The function is based on the multivariate means of the two groups (their cloud centroids in multivariate space) and their covariance matrices (the clouds' dispersions or "shapes" in multivariate space). The function is located such that the distance between the group data clouds is maximized while, simultaneously the clouds' dispersion is minimized. After calculation of the function, multivariate observations of unknown assignment may be classified as to membership of one or other of the two groups.



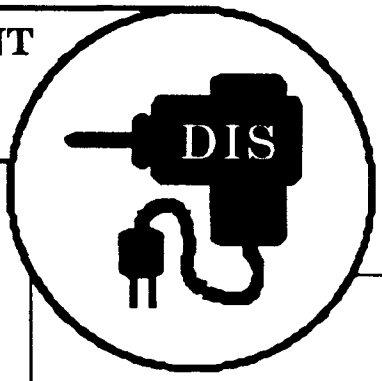
Discriminant function : $Z = \lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots \lambda_m X_m$

The MAHALANOBIS DISTANCE, D is the generalized distance between the two groups : $D^2 = \lambda_1(\bar{X}1_1 - \bar{X}2_1) + \ldots \lambda_m(\bar{X}1_m - \bar{X}2_m)$

An F ratio is calculated from : $$F = \left[\frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2)m}\right]\left[\frac{n_1 n_2}{n_1 + n_2}\right]D^2$$

and the null hypothesis that the two group centroids are equal is evaluated as an F- test with m and (n + n - m -1) degrees of freedom. If judged significant, the function can be used for classification.

DISCRIMINANT FUNCTION ANALYSIS



INPUT:
Two groups
Group interval depth ranges
Log variables

COMPUTES: Discriminant function

$$Z = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \ldots\ldots$$

Fine Print:
Parametric method
Normal distribution
Groups have equal variance / covariance matrices.

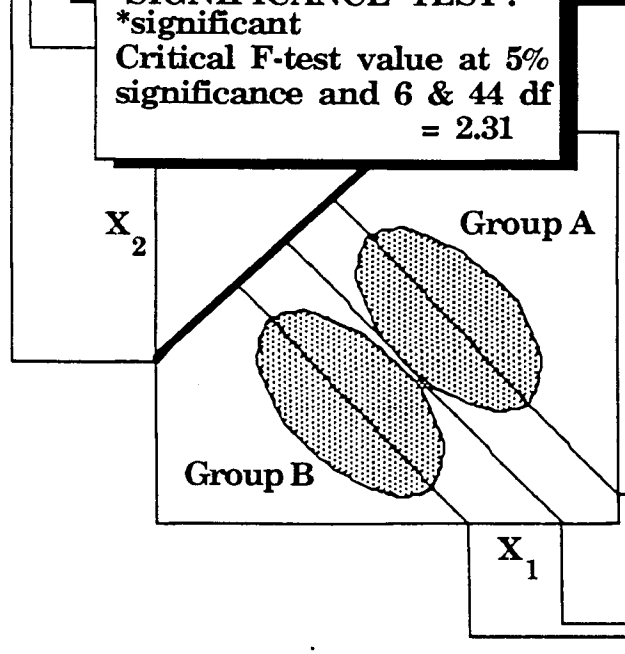
```
WELL NAME:  CHASE
LOCATION:
DATE:

GROUP A      29 OBSERVATIONS
STARTING DEPTH  2786.000        ENDING DEPTH  2800.000

GROUP B      22 OBSERVATIONS
STARTING DEPTH  2800.000        ENDING DEPTH  2811.000
```

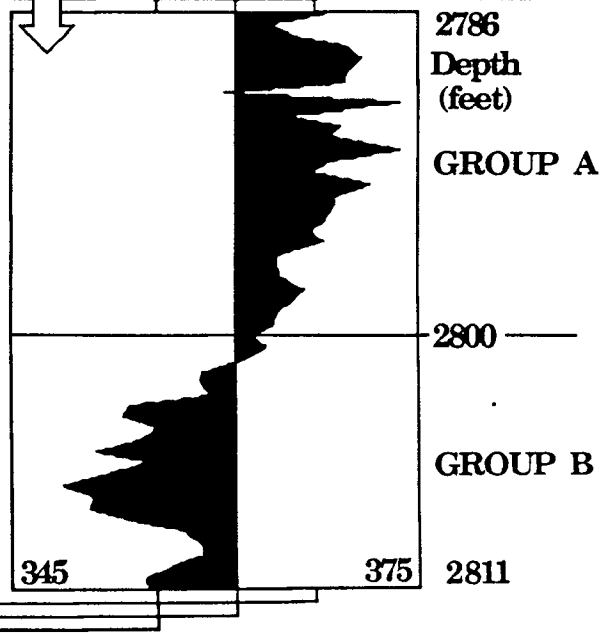
	VECTOR MEAN OF GROUP A	VECTOR MEAN OF GROUP B	DISCRIMINANT COEFFICIENT	RELATIVE CONTRIBUTION
RHOMAA	2.798	2.743	131.896	.611
UMAA	8.907	8.170	-1.987	-.125
CNL%	12.656	13.235	.330	-.016
TH	1.512	.812	5.915	.354
UR	1.050	1.506	-.579	.022
K	.606	.295	5.750	.153

```
DISCRIMINANT INDEX
     GROUP A        367.3459
     TOTAL GROUP    361.4985
     GROUP B        355.6523

STATISTICS
     MAHALANOBIS DISTANCE    11.6937
     T-TEST                 146.2863
     F-TEST                  21.8931
     DEGREES OF FREEDOM      6 AND     44
COMPUTE DISCRIMINANT SCORES - YES OR NO ....
```

Group multivariate means

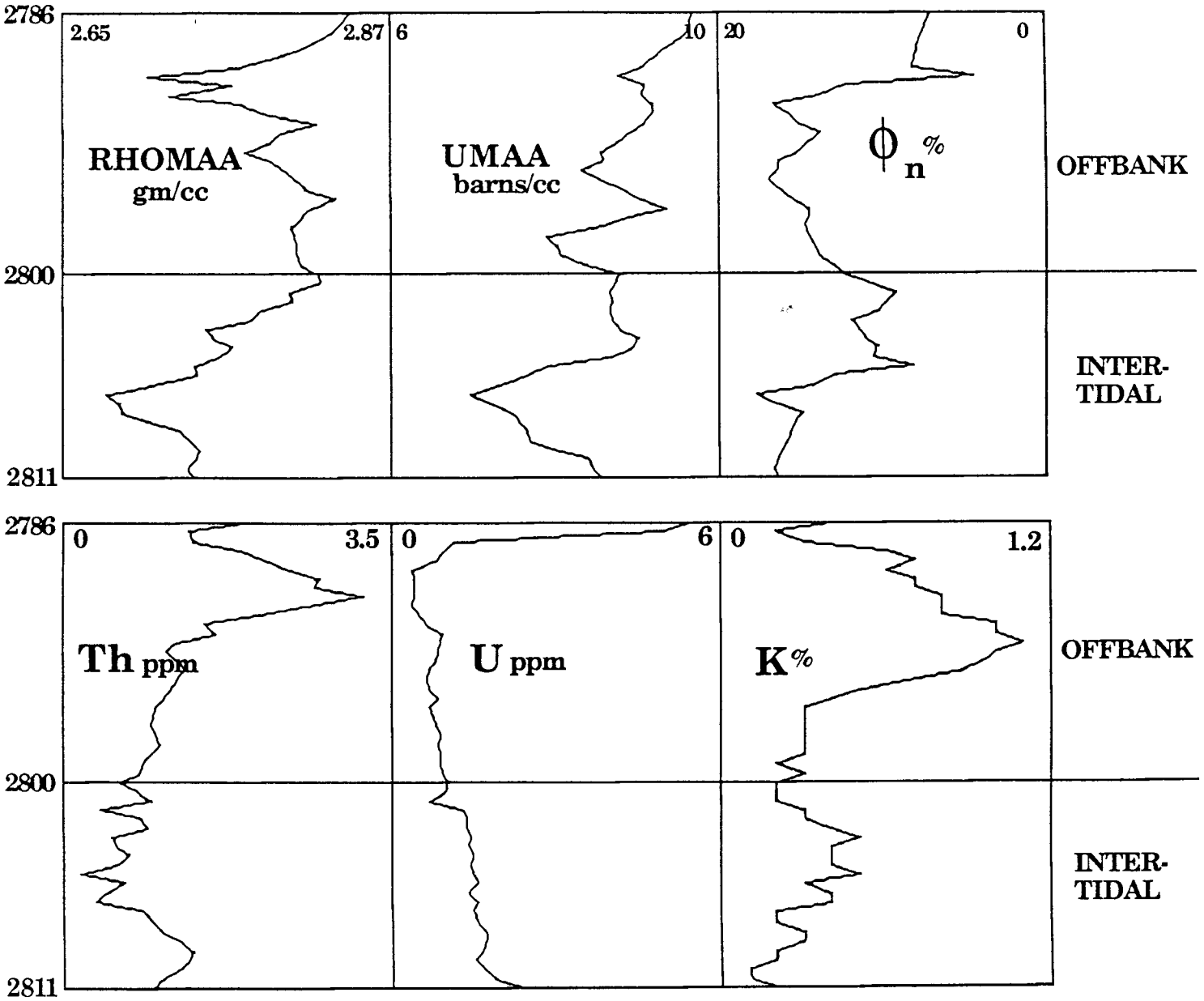
SIGNIFICANCE TEST:
*significant
Critical F-test value at 5% significance and 6 & 44 df = 2.31

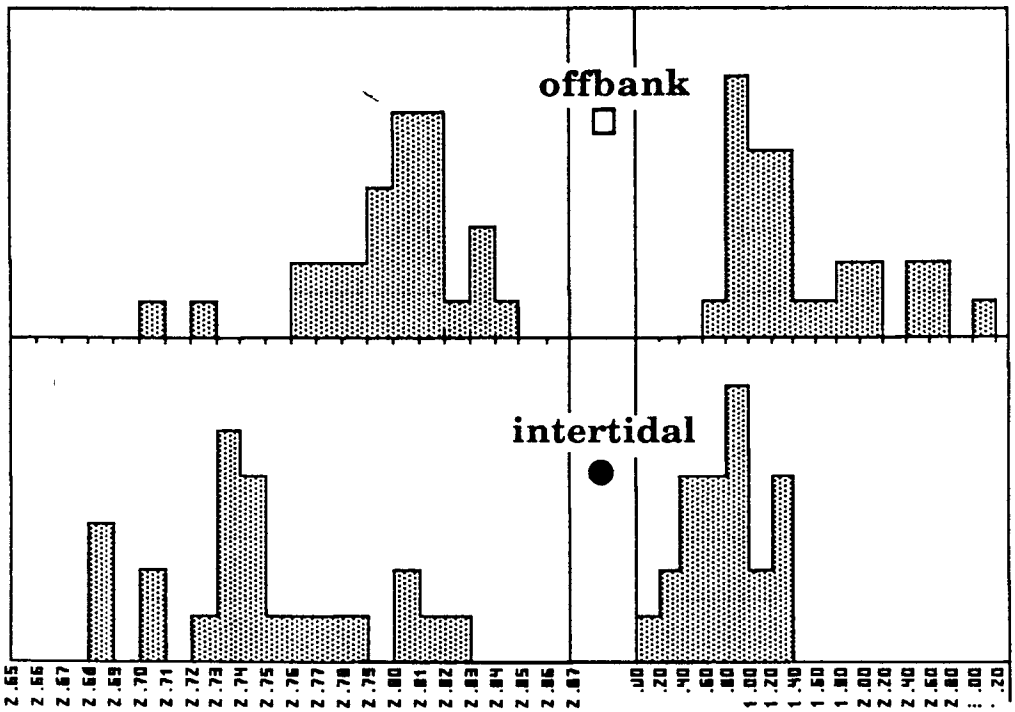


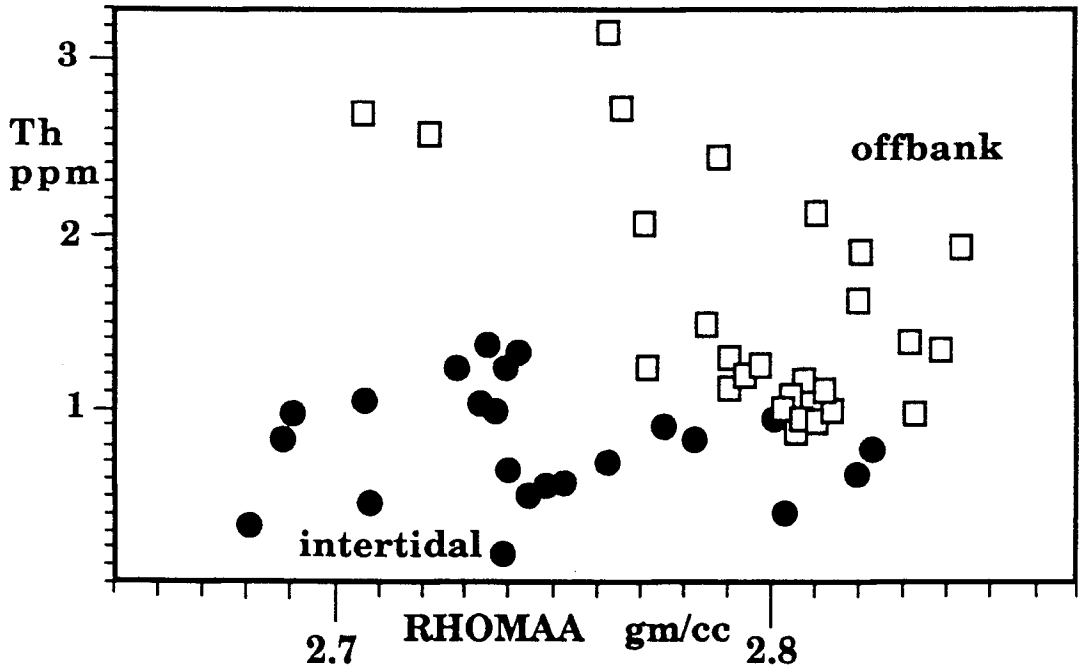


DISCRIMINANT FUNCTION ANALYSIS

Example: Distinction of Offbank and Intertidal carbonate facies in the Winfield Limestone (Chase Group) based on 6 log variables

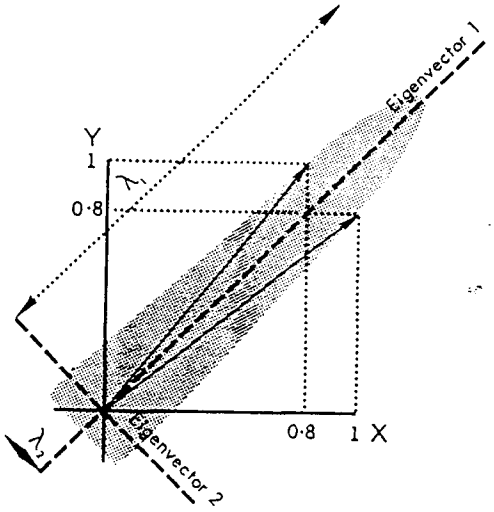



offbank
intertidal
RHOMAA gm/cc
Th ppm


Th ppm
3
2
1
offbank
intertidal
2.7
RHOMAA gm/cc
2.8

PRINCIPAL COMPONENT ANALYSIS

An UNSUPERVISED multivariate statistical technique. The principal components are the EIGENVECTORS of the covariance or correlation (standardized covariance) matrix. The eigenvectors are the principal axes of the natural variation of the data cloud which is represented by the covariance matrix as a hyperellipsoid whose centroid is the coordinates of the multivariate mean. These principal components are ordered in importance according to their EIGENVALUES which represent their relative degree of "stretch". Each eigenvalue divided by their total sum is the proportion of the total variation accounted for by the associated principal component.

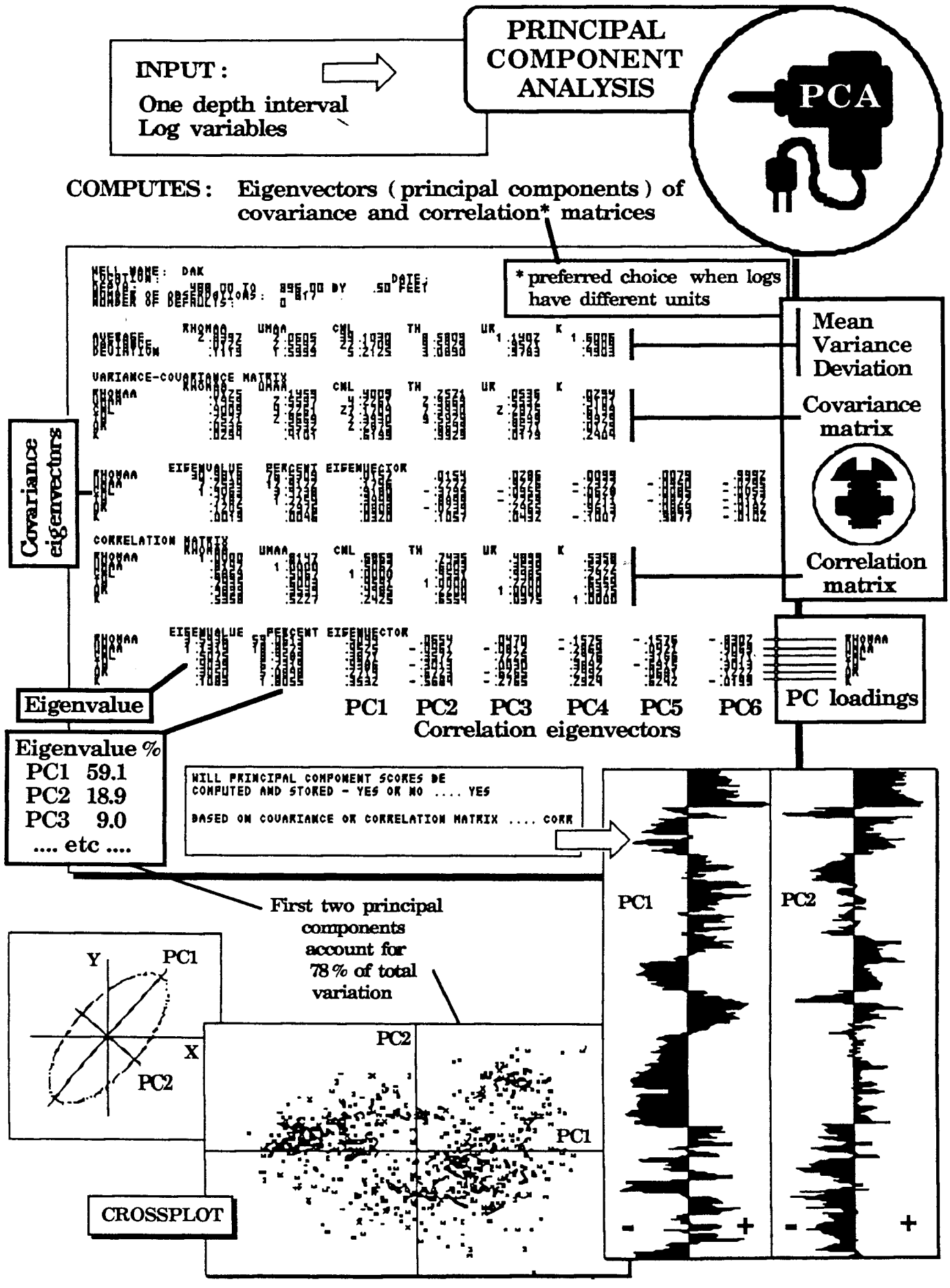


One of the main applications of principal component analysis is to reduce the unwieldy dimensionality of a data set with a large number of variables to a smaller number of dimensions. Principal components does this condensation with the minimum "damage" to the information content, because the principal components "soak up" the major sources of variation from largest to smallest. The majority of variation within data sets with many variables can often be mapped in two dimensions. This is because of the information redundancy in the data set implied by correlations between the variables.

The location of the pth principal component, Y_p with respect to the original variable axes, X_m is given by the coefficients , a_m, so that:

$$Y_p = a_1X_1 + \cdots\cdots a_mX_m$$

The coefficients can often be "read" as clues to the "meaning" of the principal component as an intrinsic process variable. The original obsevations can be converted into PRINCIPAL COMPONENT SCORES, which can be crossplotted or used as new variables in their own right.


PRINCIPAL COMPONENT ANALYSIS
PCA
INPUT :
One depth interval
Log variables
COMPUTES : Eigenvectors (principal components) of covariance and correlation* matrices
* preferred choice when logs have different units
Mean
Variance
Deviation
Covariance matrix
Correlation matrix
Covariance eigenvectors
Eigenvalue
PC loadings
PC1 PC2 PC3 PC4 PC5 PC6
Correlation eigenvectors
Eigenvalue %
PC1 59.1
PC2 18.9
PC3 9.0
.... etc ....
WILL PRINCIPAL COMPONENT SCORES BE
COMPUTED AND STORED - YES OR NO .... YES
BASED ON COVARIANCE OR CORRELATION MATRIX .... CORR
First two principal components account for 78 % of total variation
Y
PC1
X
PC2
PC2
PC1
CROSSPLOT
PC1
PC2
-
+
-
+

PRINCIPAL COMPONENT ANALYSIS OF LOGS IN LOWER CRETACEOUS SANDSTONES AND SHALES

Input log variables :

- RHOMAA
- UMAA
- CNL neutron
- Thorium
- Uranium
- Potassium

Input logs are standardized to eliminate influence of different measurement units. The variance - covariance matrix is then the correlation matrix.
The eigenvectors of the correlation matrix are the principal components of the standardized logs, and the eigenvalues express the relative amount of the total variation accounted for by each principal component.

Correlation matrix

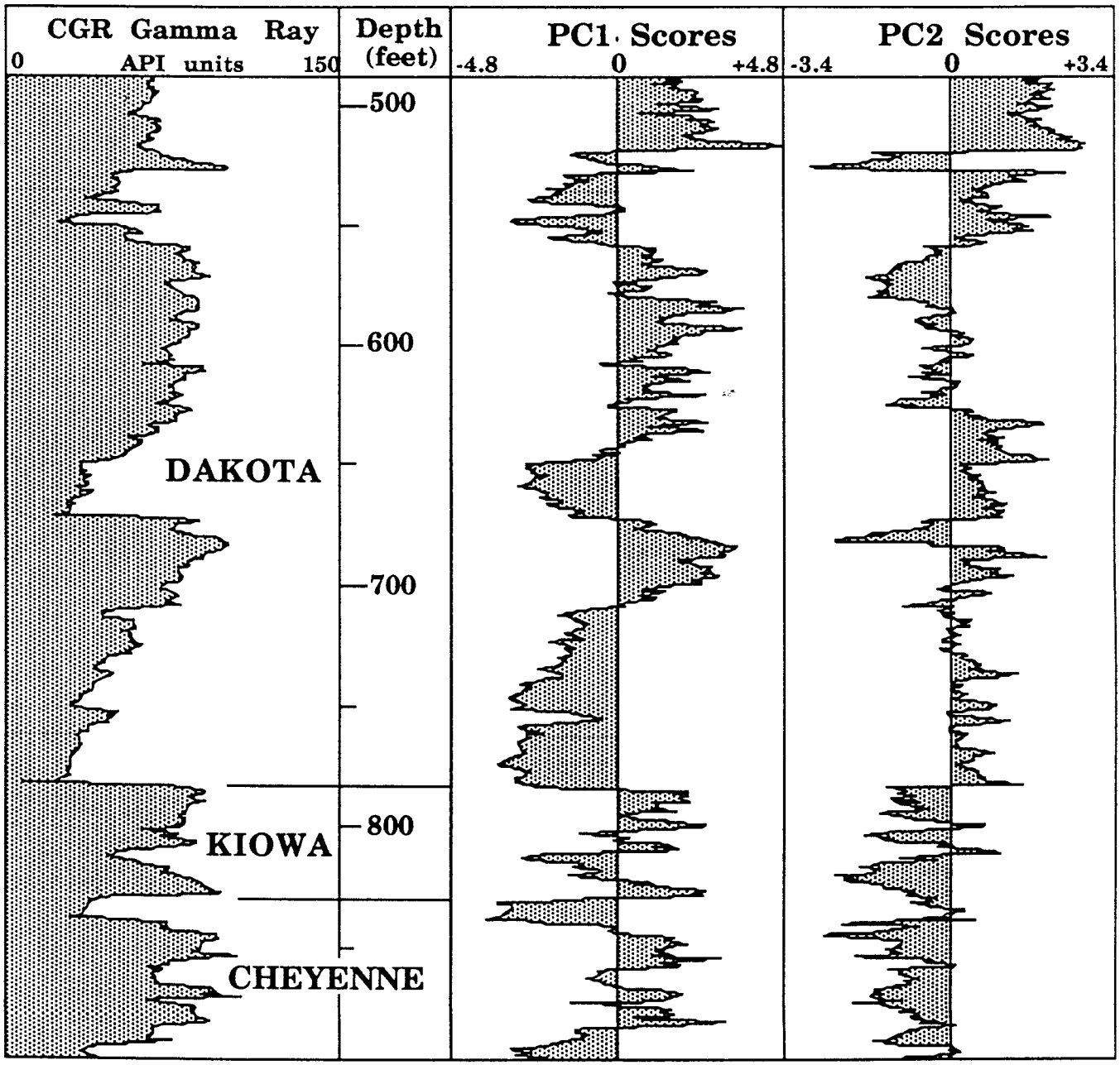
	RHOMAA	UMAA	CNL	TH	UR	K
RHOMAA	1.0000	.8147	.6869	.7435	.4899	.5358
UMAA	.8147	1.0000	.5067	.6003	.3534	.5227
CNL	.6869	.5067	1.0000	.4591	.4485	.2425
TH	.7435	.6003	.4591	1.0000	.2200	.6554
UR	.4899	.3534	.4485	.2200	1.0000	.0375
K	.5358	.5227	.2425	.6554	.0375	1.0000

Principal component eigenvectors

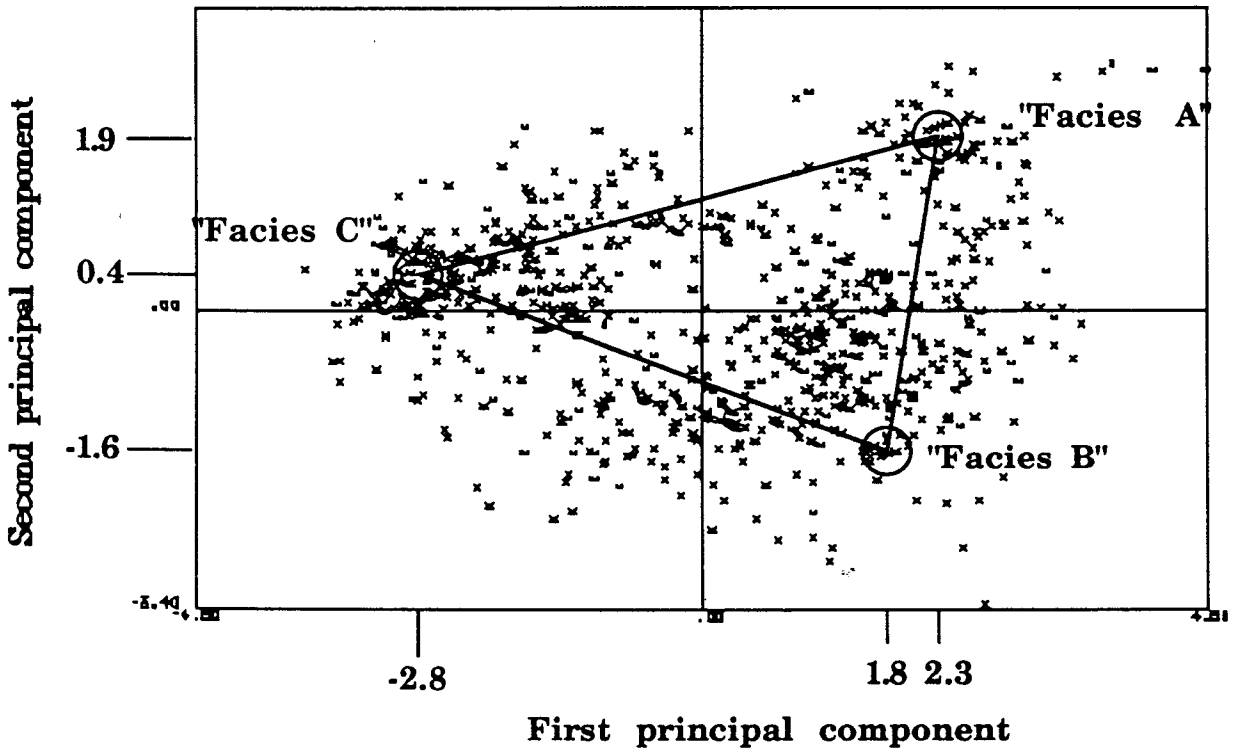
	PC1	PC2	PC3	PC4	PC5	PC6
RHOMAA	.5044	.0654	.0470	-.1575	-.	
UMAA	.4525	-.0461	-.0817	-.7864	.	
CNL	.3871	.3572	.7225	.2476	.	Etc.
TH	.4386	-.3013	.0030	.3842	-.	
UR	.2714	.6723	-.6265	.2492	.	
K	.3532	-.5684	-.2765	.2924	.	
Eigenvalue %	59%	19%	8%	7%	5%	2%

principal component scores

PRINCIPAL COMPONENT SCORE LOGS



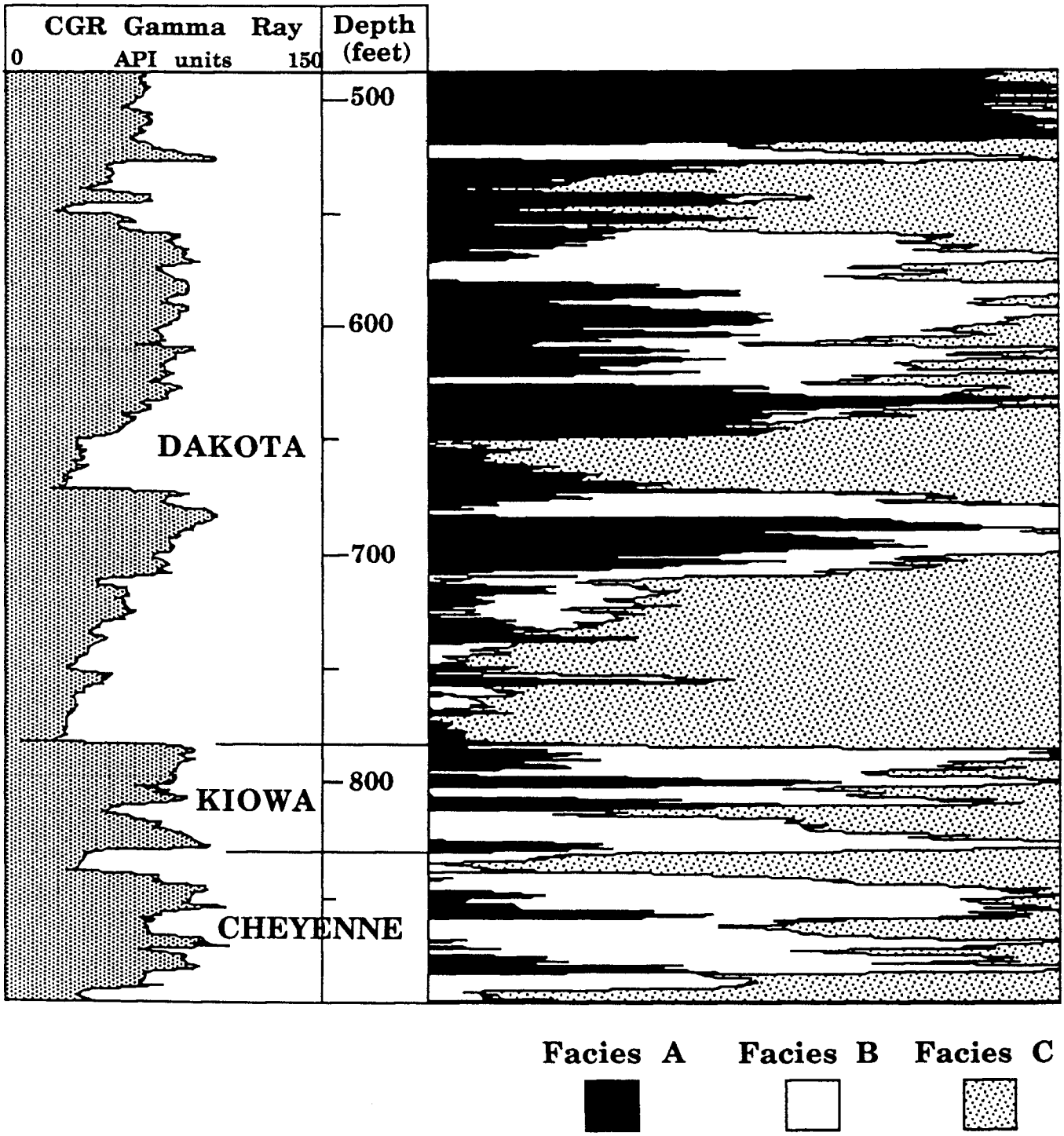
PATTERN RECOGNITION OF ELECTROFACIES ON CROSSPLOT OF FIRST TWO PRINCIPAL COMPONENTS



PC score coefficients for compositional analysis by matrix algebra solution :

	PC1	PC2
Electrofacies A	2.3	1.9
B	1.8	-1.6
C	-2.8	0.4

Matrix algebra solution of electrofacies located on crossplot of first and second principal components of RHOmaa, Umaa, Φn, Th, U, K



FACTOR ANALYSIS

FACTOR ANALYSIS is a theoretical model that postulates that observed variables are correlated with a lesser number of "hidden" variables or FACTORS which explain the systematic variation in the measured sample. The theory was developed in psychology, where factors were considered to be aptitudes or personality traits which accounted for patterns in results from examinations or questionaires.

The initial phase of factor analysis is programmed as a principal component solution, which generates m eigenvectors for m variables. However, while principal components are simply a geometrical result, factor analysis is a model, in which a FEWER causal variables are said to explain the data. Traditionally, the number and identity of the factors are supposed to be known beforehand and dictate the design of the exam or poll. This is usually not the case in geology. However, the PCA solution of eigenvalues can give insight on the intrinsic dimensionality of the data, and so the number of factors.

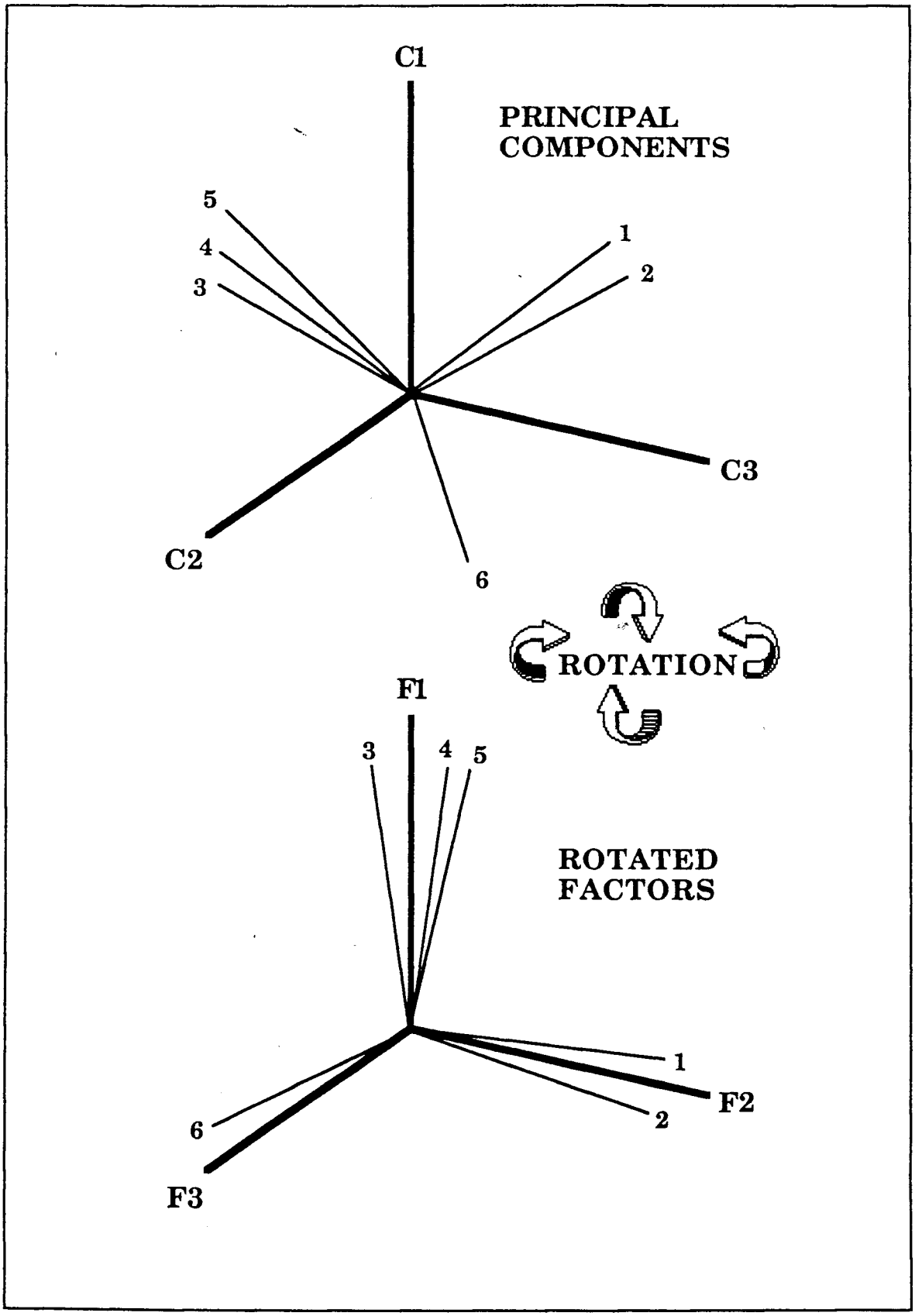
The factor model for k factors can be written as :

$$Z_j = a_{1j}F_1 + a_{2j}F_2 + \ldots + a_{kj}F_k + a_jE_j$$

where Z_j is the jth factor, a_k are the FACTOR LOADINGS, and E_j is the residual error. This equation may be used to transform the raw variables for any observation into a FACTOR SCORE.

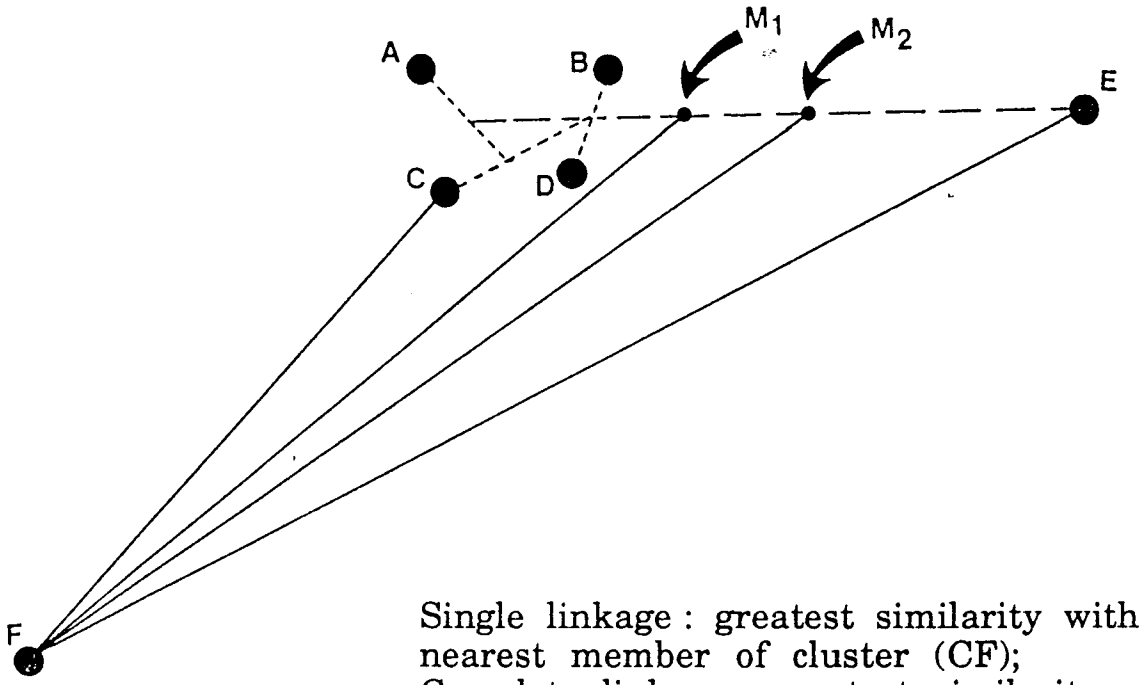
The initial principal components solution is unsatisfactory as a factor solution, both because it has as many components as the original variables, but also because the eigenvectors are generalized composites of the major sources of variation. The aim of factor analysis is to locate "simple" structure, in which each of the variables is expressed as either a strong loading (ideally, +1 or -1), or a weak loading (ideally, zero) with respect to each of the factors. The orthogonal axes are therefore rotated from the eigenvectors in search of a solution with a simple structure.

Once the factor model is solved, the factors may be interpreted as causal variables, based on the factor loadings of the original variables. The factor scores also provide new variables which, hopefully, are more closely linked with the phenomenom of interest than any of the original variables.

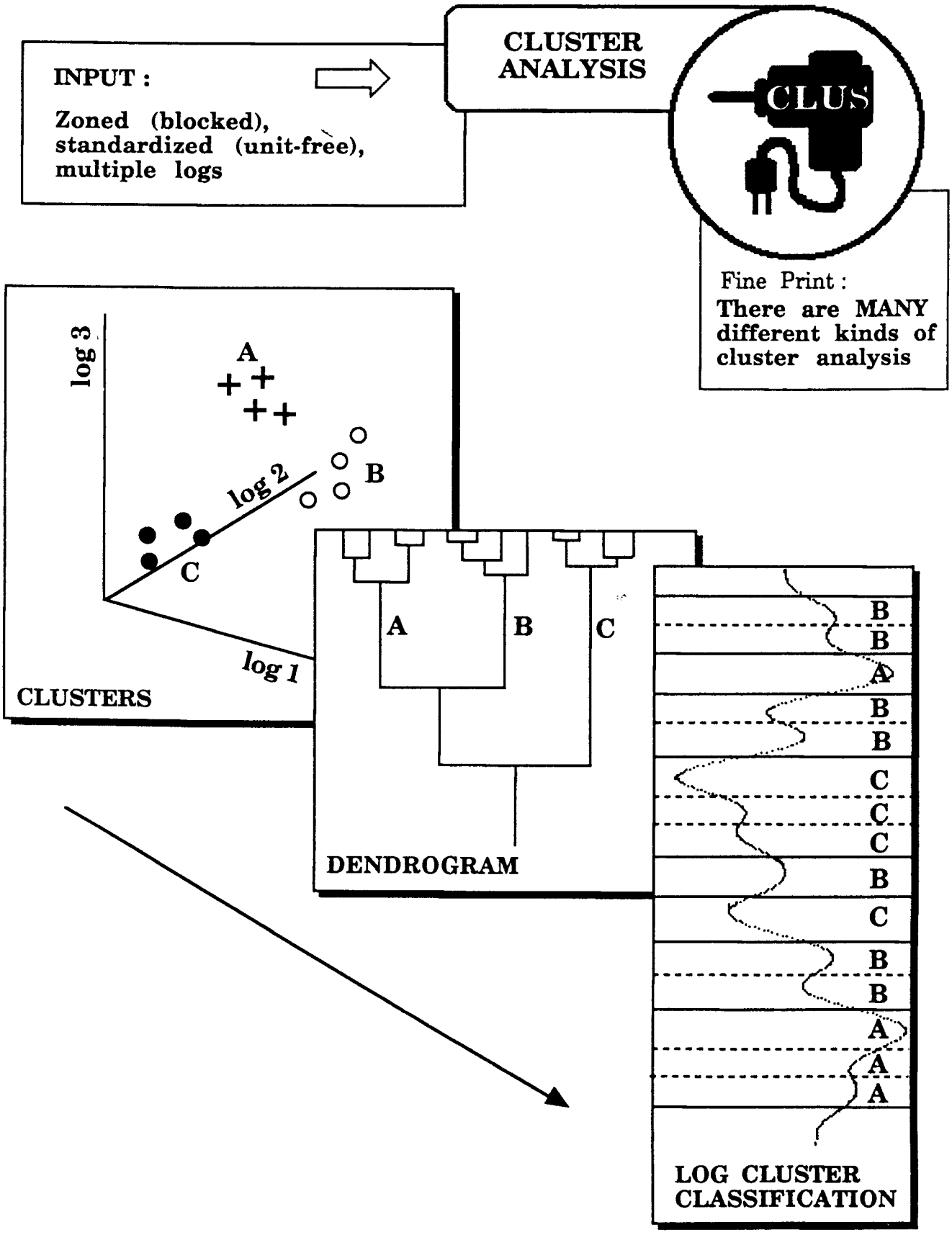

C1
PRINCIPAL
COMPONENTS
5
4
3
1
2
C3
C2
6
ROTATION
F1
3
4
5
ROTATED
FACTORS
1
F2
2
6
F3

CLUSTER ANALYSIS

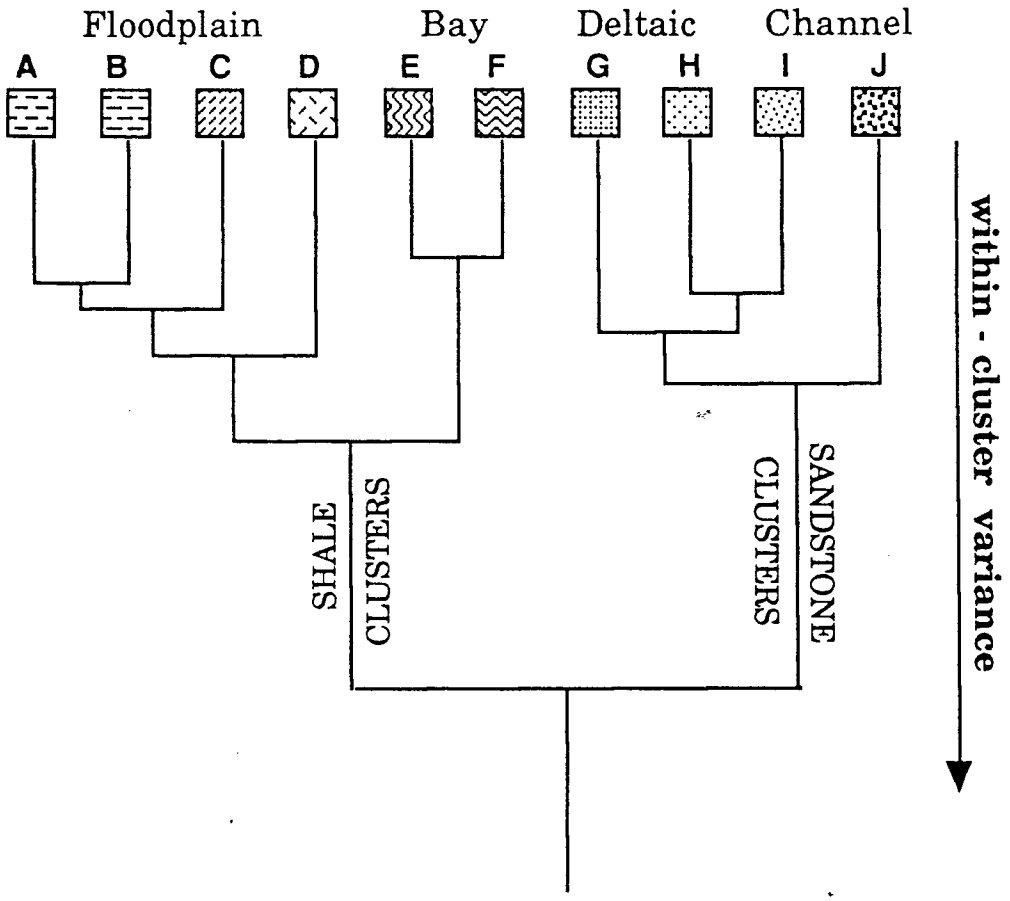
There are a great variety of CLUSTER ANALYSIS methods, each based on different strategies to assign observations to groups which are "homogeneous" and distinct from other groups. The most common techniques used in geology are those of HIERARCHIC CLUSTERING, largely because of their popularity as methods for NUMERICAL TAXONOMY of fossils. First an n x n matrix of similarities between all pairs of the n observations is calculated. Then the pairs with the highest similarities are merged, the matrix recomputed, and the procedure repeated until some "natural" level of clustering is reached, as assessed by visual inspection or some numerical criterion. The result is most commonly shown as a DENDROGRAM. Even in hierarchical clustering, there are different strategies to link observations with clusters in formation as illustrated below :



Single linkage : greatest similarity with nearest member of cluster (CF);
Complete linkage : greatest similarity with most dissimilar member of cluster (EF);
Average linkage : greatest average similarity with members of cluster (M2F);
Centroid linkage : least distance to centroid of cluster (M1F).

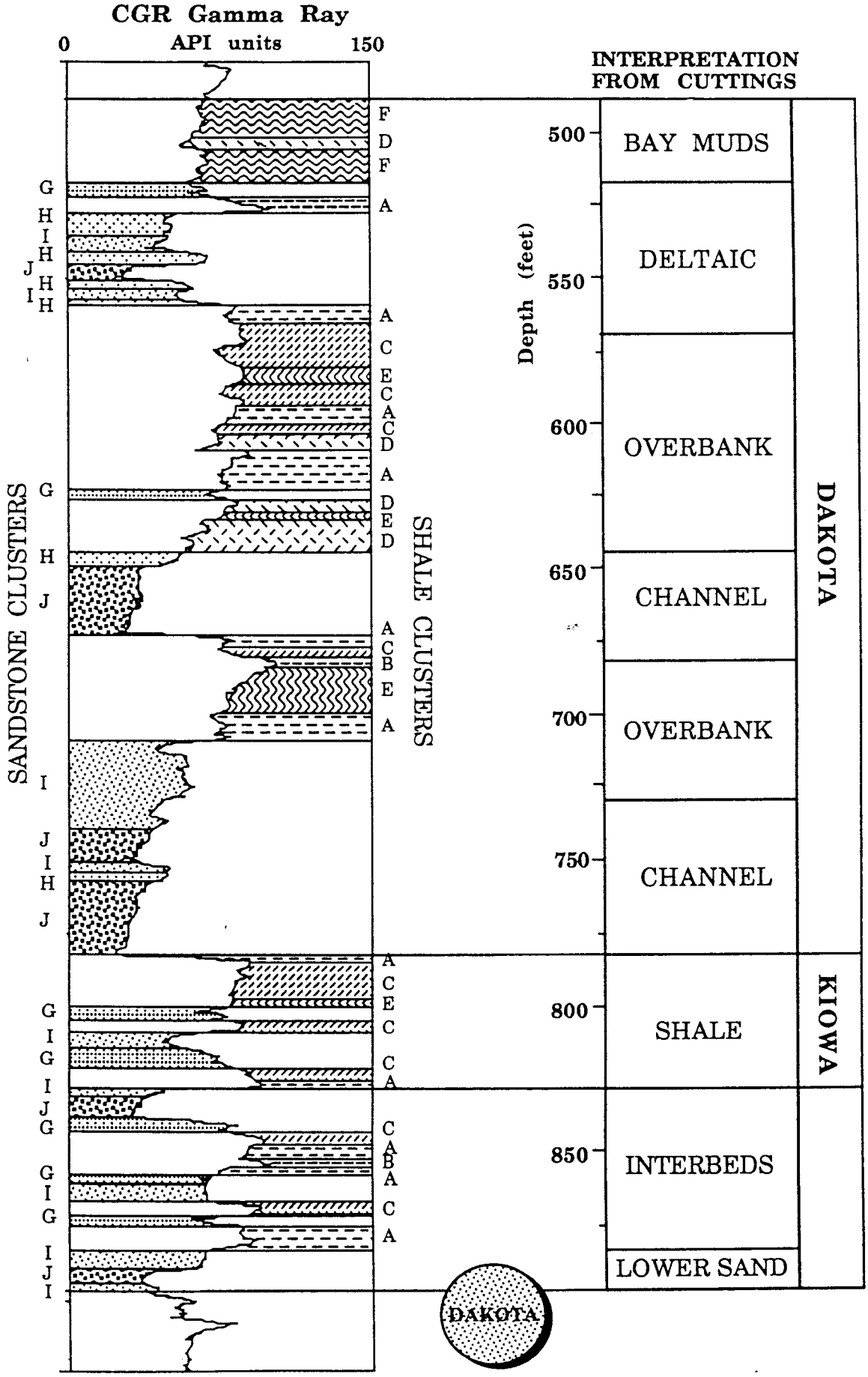

INPUT :
Zoned (blocked), standardized (unit-free), multiple logs
CLUSTER ANALYSIS
CLUS
Fine Print :
There are MANY different kinds of cluster analysis
log 3
log 2
log 1
A
B
C
CLUSTERS
A
B
C
DENDROGRAM
B
B
A
B
B
C
C
C
B
C
B
B
A
A
A
LOG CLUSTER CLASSIFICATION

CLUSTER ANALYSIS OF ZONES IN A SANDSTONE - SHALE SEQUENCE



Input log variables for clustering :

Apparent grain density (RHOmaa)
Apparent matrix photoelectric cross-section (Umaa)
Neutron porosity (ϕ_n)
Potassium (K)
Uranium (U)
Thorium (Th)


CGR Gamma Ray
0
API units
150
INTERPRETATION FROM CUTTINGS
SANDSTONE CLUSTERS
SHALE CLUSTERS
Depth (feet)
500
550
600
650
700
750
800
850
BAY MUDS
DELTAIC
OVERBANK
CHANNEL
OVERBANK
CHANNEL
SHALE
INTERBEDS
LOWER SAND
DAKOTA
KIOWA
DAKOTA

CASE STUDIES

Hiawatha Designs an Experiment

Maurice G. Kendall

1. Hiawatha, mighty hunter
   He could shoot ten arrows upwards
   Shoot them with such strength and swiftness
   That the last had left the bowstring
   Ere the first to earth descended.
   This was commonly regarded
   As a feat of skill and cunning.

2. One or two sarcastic spirits
   Pointed out to him, however,
   That it might be much more useful
   If he sometimes hit the target.
   Why not shoot a little straighter
   And employ a smaller sample?

3. Hiawatha, who at college
   Majored in applied statistics
   Consequently felt entitled
   To instruct his fellow men on
   Any subject whatsoever,
   Waxed exceedingly indignant
   Talked about the law of error,
   Talked about truncated normals,
   Talked of loss of information,
   Talked about his lack of bias
   Pointed out that in the long run
   Independent observations
   Even though they missed the target
   Had an average point of impact
   Very near the spot he aimed at
   (With the possible exception
   Of a set of measure zero.)

4. This, they said, was rather doubtful.
   Anyway, it didn't matter
   What resulted in the long run;
   Either he must hit the target
   Much more often than at present
   Or himself would have to pay for
   All the arrows that he wasted.

5. Hiawatha, in a temper
   Quoted parts of R. A. Fisher
   Quoted Yates and quoted Finney
   Quoted yards of Oscar Kempthorne
   Quoted reams of Cox and Cochran
   Quoted Anderson and Bancroft
   Practically in extenso
   Trying to impress upon them
   That what actually mattered
   Was to estimate the error.

6. One or two of them admitted
   Such a thing might have its uses
   Still, they said, he might do better
   If he shot a little straighter.

7. Hiawatha, to convince them
   Organized a shooting contest
   Laid out in the proper manner
   Of designs experimental
   Recommended in the textbooks
   (Mainly used for tasting tea, but

Reprinted from *American Statistician*, 1959, 13, 23–24, by permission of the publisher.

Sometimes used in other cases)
Randomized his shooting order
In factorial arrangements
Used in the theory of Galois
Fields of ideal polynomials
Got a nicely balanced layout
And successfully confounded
Second-order interactions.

8. All the other tribal marksmen
Ignorant, benighted creatures,
Of experimental set-ups
Spent their time of preparation
Putting in a lot of practice
Merely shooting at a target.

9. Thus it happened in the contest
That their scores were most impressive
With one solitary exception
This (I hate to have to say it)
Was the score of Hiawatha,
Who, as usual, shot his arrows
Shot them with great strength and swiftness
Managing to be unbiased
Not, however, with his salvo
Managing to hit the target.

10. There, they said to Hiawatha,
That is what we all expected.

11. Hiawatha, nothing daunted,
Called for pen and called for paper
Did analyses of variance
Finally produced the figures
Showing beyond peradventure
Everybody else was biased
And the variance components
Did not differ from each other
Or from Hiawatha's
(This last point, one should acknowledge
Might have been much more convincing
If he hadn't been compelled to
Estimate has own component
From experimental plots in
Which the values all were missing.
Still, they didn't understand it
So they couldn't raise objections
This is what so often happens
With analyses of variance).

12. All the same, his fellow tribesmen
Ignorant, benighted heathens,
Took away his bow and arrows,
Said that though my Hiawatha
Was a brilliant statistician
He was useless as a bowman,
As for variance components
Several of the more outspoken
Made primeval observations
Hurtful to the finer feelings
Even of a statistician.

13. In a corner of the forest
Dwells alone my Hiawatha
Permanently cogitating
On the normal law of error.
Wondering in idle moments
Whether an increased precision
Might perhaps be rather better
Even at the risk of bias
If thereby one, now and then, could
Register upon the target.

REGIONAL POROSITY VARIATIONS IN THE LEDUC REEFS: APPLICATION OF QUANTILES AND BOX PLOTS

Amthor et al (1994) studied porosity and permeability patterns in the Upper Devonian Leduc reservoirs of the Rimbey-Meadowbrook trend of southern Alberta, Canada (Figs. 1 and 2). They found that, if no account was made of burial depth, limestones and dolomitic limestones were more porous than dolomites. Moving southwards down the trend (Fig.1) the reservoirs are located at increasing depths of burial and there is a systematic decrease in porosity. This trend is shown well by the plot of porosity quantiles versus township in Figure 3. Porosity measurements are commonly (but not invariably) satisfactorily approximated by a normal distribution provided that they are not multimodal. In these cases, trends of this type would be shown adequately by plots of means and standard deviation ranges. However, many of the Leduc porosities are sufficiently low, so that, even if normal, their distributions would be truncated-normal, with a resulting asymmetry. Therefore, the use of quantiles in this example (the lower 10th, the median, and the higher 90th) is a good choice as can be seen on the plot of Figure 3.

Amthor et al (1994) also used box plots to demonstrate the differences between the porosities and permeabilities of different carbonate lithologies, carbonates of reef buildups and platform, and porosities contrasted by depth. Box plots are a good graphic medium for their thesis, because their simplicity allows readers to follow multiple comparisons of different porosity groups with relative ease. The box plots of Figure 4 show comparison between porosities in a shallow limestone reservoir (Golden Spike), a shallow dolomite reservoir (Leduc), and deep limestone and dolomite reservoir sections (Strachan). The box plots show that, while they confirm the overall decrease in porosity with depth, they also reflect the trend that limestones lose most of their porosity with depth, but that dolomites tend to retain much of their porosity.

Amthor et al (1994) point out that these conclusions have economic significance for exploration within the deep Alberta basin. Dolomitized buildups should be favored over limestones. However, the statistics could also be used in more detailed analysis for pre-drill predictions of porosity (and permeability) ranges of Leduc reef prospects.

REFERENCE

Amthor, J.E., Mountjoy, E.W., and Machel, H.G., 1994, Regional-scale porosity and permeability variations in Upper Devonian Leduc buildups: implications for reservoir development and prediction in carbonates: AAPG Bulletin, v.78, no. 10, p. 1541-1569.

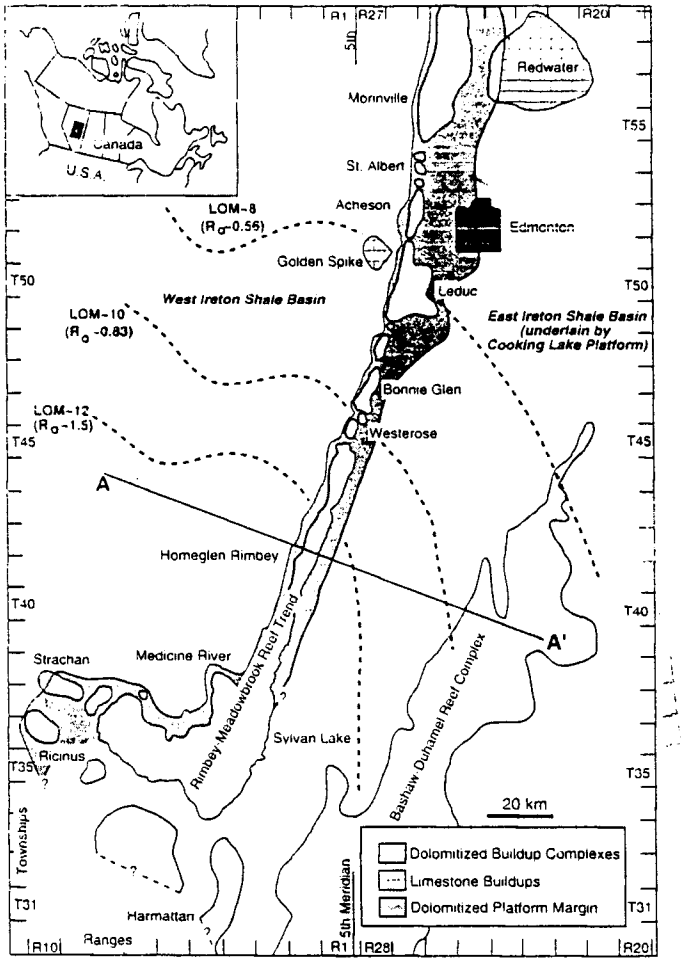
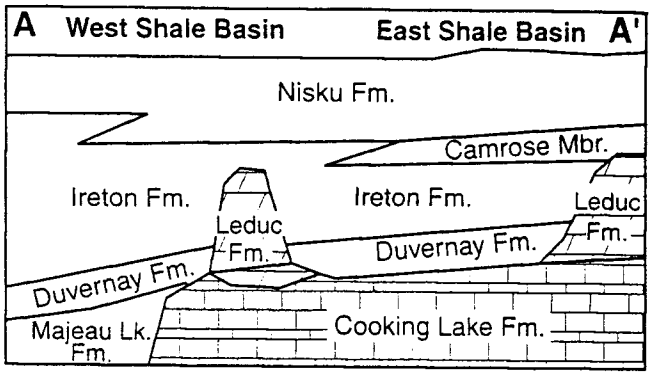
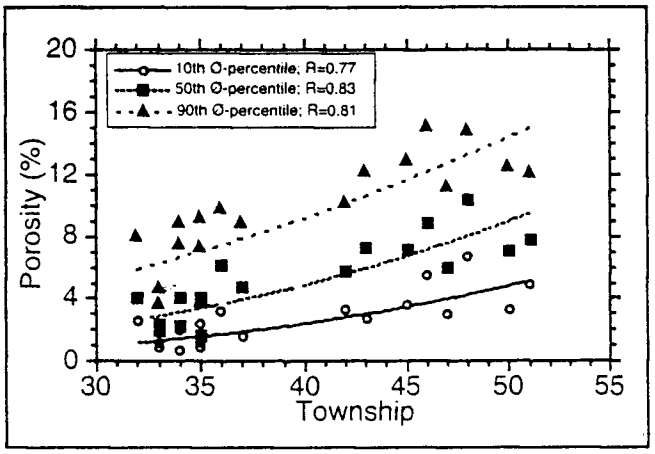


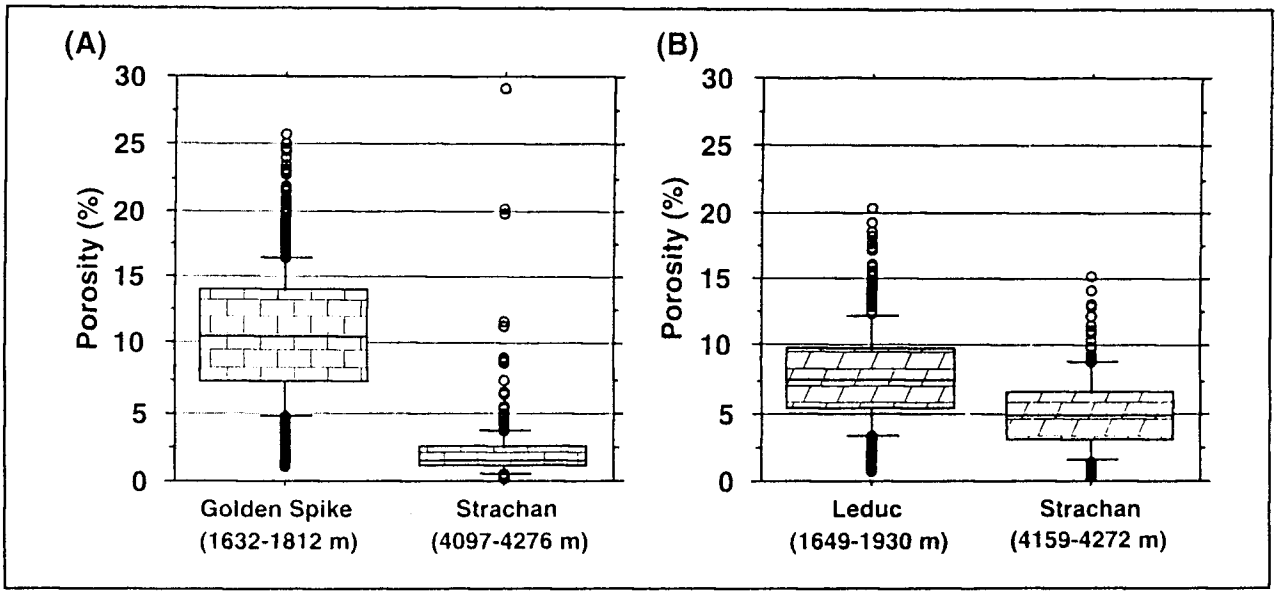
Figure 1—Map of the Rimbey-Meadowbrook reef trend, showing buildups and extent of dolomitization in the Cooking Lake carbonate platform (dark stippled pattern). Buildups are dolomitized where the margin of the Cooking Lake platform is dolomitized. Golden Spike and Redwater are situated off the margin and are not dolomitized.



–Schematic Devonian subsurface stratigraphy of the study area (after Stoakes, 1980). For approximate line of section AA' see Figure 1.



—Well-scale porosity vs. geographic location (Township, TWP) for 18 wells of the Rimbey-Meadowbrook reef trend (see Figure 1 for location of townships). Data points represent low (10th porosity percentile), median (50th porosity percentile), and high (90th porosity percentile) porosity values of a single well.



–Boxplots showing porosity distributions in limestone (A) and dolomitized buildups (B) at different burial depths. Limestones of the Golden Spike buildup (A) show higher porosity values than dolostones of the adjacent Leduc buildup (B) at shallow burial depths (<2000 m). At burial depths greater than 4000 m, this relationship is reversed: limestones have lost most of their porosity (e.g., Strachan buildup in A), whereas dolostones retain more of their porosity (e.g., Strachan buildup in B).

CHARACTERIZATION OF AVERAGE PERMEABILITY IN TIGHT GAS FORMATIONS: COMPARISON OF THE ARITHMETIC AND GEOMETRIC MEANS AND THE MEDIAN

The 1978 Natural Gas Policy Act (NGPA) gives financial incentives to stimulate gas production from low-permeability formations. In order to qualify, a formation pay section must be shown to have an average permeability of less than 0.1 md. The regulations do not specify how the average should be estimated. In Texas, the Cleveland, Cotton Valley, Lobo, and Travis Peak formations have all been considered for classification as tight gas formations under the NGPA rules. Rollins et al (1992) collected permeability data from the four formations, and found their distributions to be unimodal and positive-skewed similar to a lognormal distribution (see Figures of Travis Peak formation permeabilities). They computed summary statistics of:

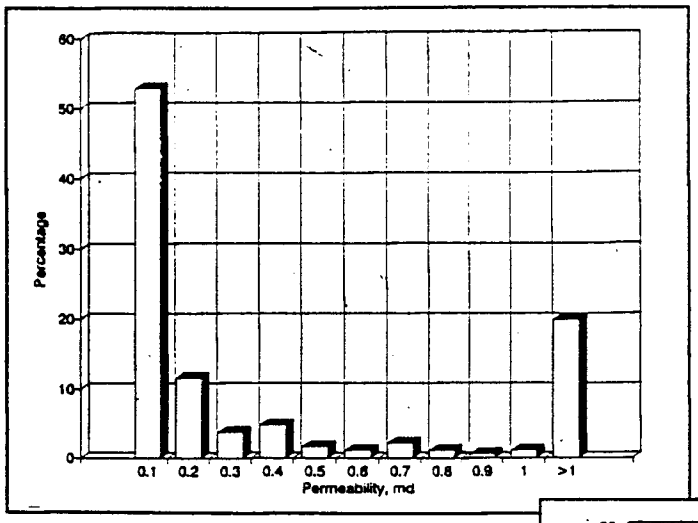
Formation	Number of wells	Arithmetic mean, md	Median, md
Cleveland	391	0.179	0.028
Cotton Valley	395	7.378	0.045
Lobo	112	0.235	0.056
Travis Peak	191	1.035	0.085

The use of the arithmetic average would disqualify all four formations; the median would designate all four as tight gas formations. In reality, large acreage blocks were selectively excluded, so that the permeability of the remaining acreage would have an arithmetic average of less than 0.1 md.

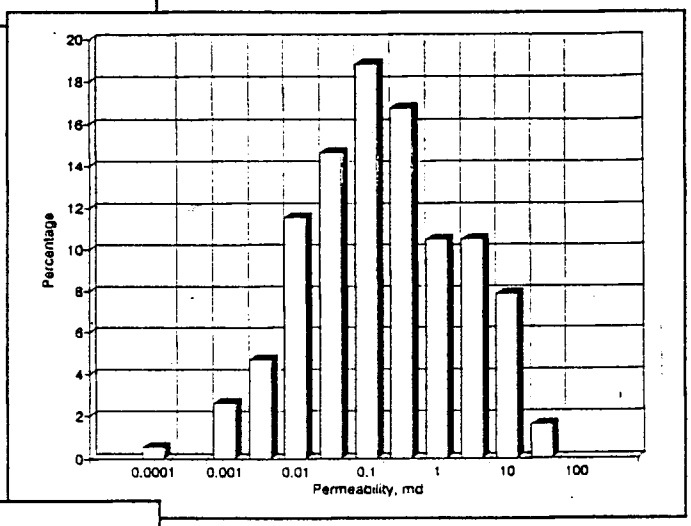
Rollins et al (1992) made a comparative study of the performance of the arithmetic mean, the geometric mean, and the median as estimates of average formation permeability through simulations of gas production in two families of wells from the Travis Peak formation. The results were:

	CASE 1	CASE 2
Arithmetic mean	1.350 md -- 5.64 Bcf	0.284 md -- 5.30 Bcf/well
Geometric mean	0.106 md -- 4.00 Bcf	0.017 md -- 1.26 Bcf/well
Median	0.085 md -- 3.62 Bcf	0.020 md -- 1.41 Bcf/well
ACTUAL	**3.42 Bcf**	**1.75 Bcf/well**

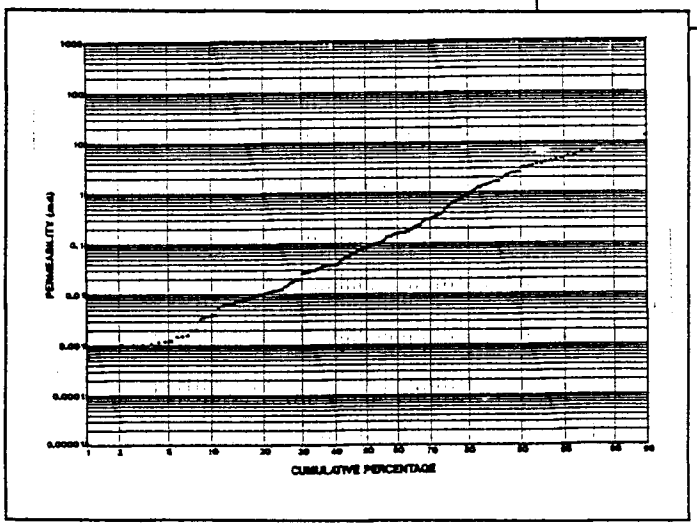
In each case, the median provided the best match and suggests that it is the "natural" average and should be the central measure statistic to be applied to NGPA qualification. If the permeability data were closely matched by the lognormal distribution, then a good theoretical case could be made for the geometric average. In fact, it performed quite well in the simulation comparisons, but tended to overestimate in one case, and underestimate in the other. Even if the permeabilities were lognormal, this characteristic would probably still be seen because the median is a more stable estimate of the center.



Permeability distribution in Travis Peak



Log permeability distribution in Travis Peak



Lognormal probability plot of Travis Peak permeabilities

From Rollins, Holditch, and Lee (1992)

At small and moderate sample sizes, the geometric mean will be less stable because of its sensitivity to extreme values in the tails.

Rollins et al (1992) pointed out that their conclusions concerning the use of the median as an appropriate average measure of permeability applied to estimates on an areal basis. They cited studies such as by Richard et al (1987) that show that the correct way to average permeability vertically in a section with multiple layers is to use a thickness-weighted arithmetic mean. These types of estimate can then be used in the calculation of deliverability and expected stabilized flow rate from a well.

REFERENCES

Richardson, J.G., Sangree, J.B., and Sneider, R.M., 1987, Permeability distributions in reservoirs: JPT, (Oct.) p. 1197-99.

Rollins, J.B., Holditch, S.A., and Lee, W.J., 1992, Characterizing average permeability in oil and gas formations: SPE Formation Evaluation, v. 7, no. 1, p. 99-105.

NON-LINEAR REGRESSION

When the descriptive equation that links two variables is of the form : $Y = aX^b$

the trend is non-linear, but may be linearized by a logarithmic transformation : $\log Y = \log a + b \log X$

Following the transformation, the data can be analyzed by simple linear regression. However, it should be realized that the squared errors that are minimized are in logarithmic, rather than arithmetic units.

The most well-known example of this equation in petrophysics is the modified Archie equation : $F = \frac{a}{\phi^m}$

An example of the solution of the Archie equation by non-linear regression is shown on the next page.

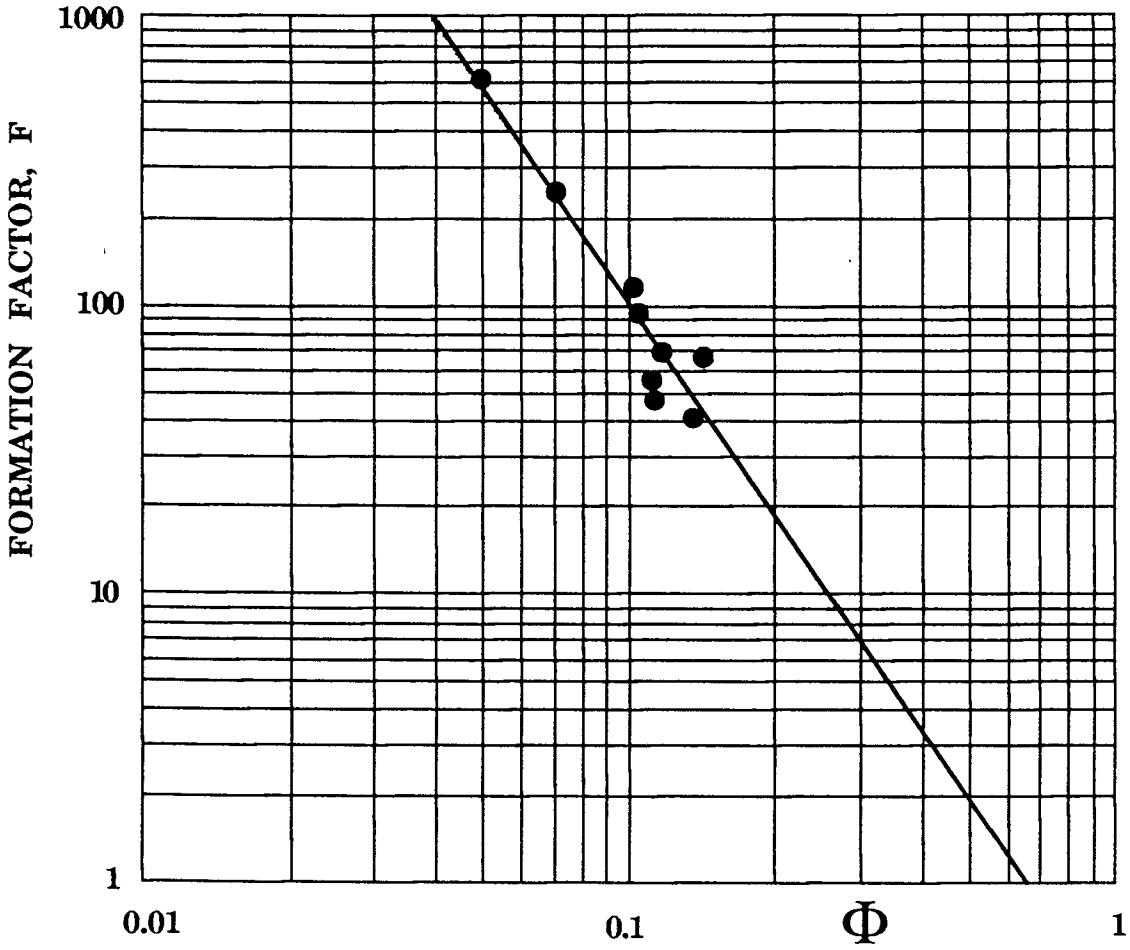
Other examples include prediction equations that link permeability and porosity such as :

$$\phi = a k^b$$ (Muskat,1949).

or the possible prediction of clay mineral exchange cations in shaly sands by some power of the porosity the equation: $Q_v = d\phi^{-e}$

as suggested by Lavers and others (1974), where the constant d, and exponent e, probably function as dimensional modifiers to convert porosity as volume to a measure of internal surface area.

EXAMPLE : REGRESSION CALCULATION OF ARCHIE EQUATION CONSTANTS FOR 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 : $F = \frac{0.27}{\Phi^{2.57}}$)

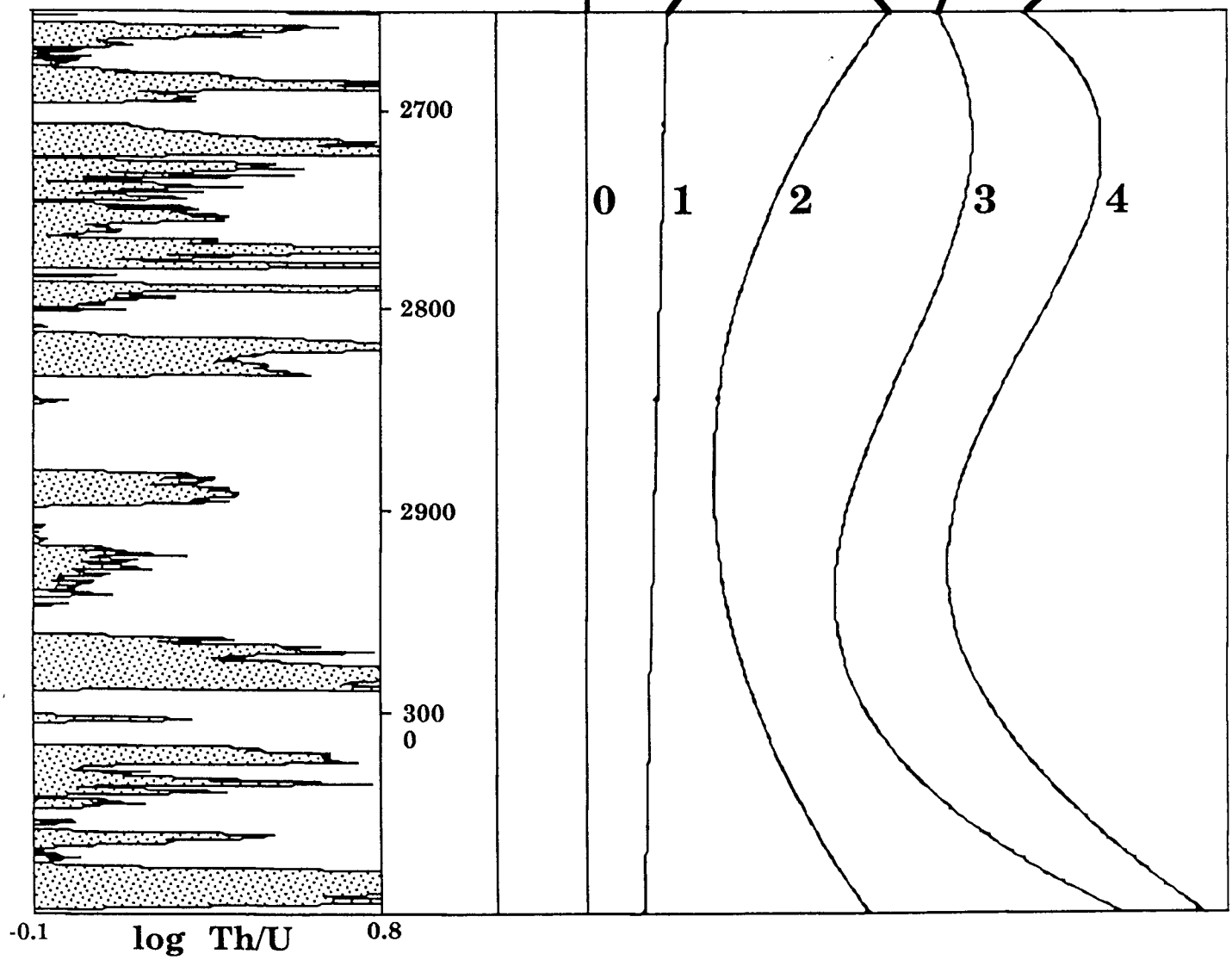
EXAMPLE : POLYNOMIAL REGRESSION OF log(Th/U) ON DEPTH IN CHASE GROUP AS POTENTIAL INDICATOR OF LONG-TERM TRENDS IN REDOX POTENTIAL

TERRALOG

WELL NAME: CHASE
LOCATION:
DATE:
DEPTH: 2650.00 TO 3100.00 BY .50 FEET
DEPENDENT VARIABLE: LOGTHU
INDEPENDENT VARIABLE: (DEP-2650)/1000

SIMPLE REGRESSION

ORDER OF REGRESSION	0	1	2	3	4
REGRESSION COEF A:	.137	.167	.451	.184	.102
REGRESSION COEF X:		-.131	-3.935	3.228	6.895
REGRESSION COEF X**2:			8.482	-31.469	-68.218
REGRESSION COEF X**3:				59.371	186.605
REGRESSION COEF X**4:					-141.738
NUMBER OF SAMPLES	895	895	895	895	895
SUM OF SQUARES TOTAL	226.7	226.7	226.7	226.7	226.7
SUM OF SQUARES REGRESSION	.0	.2	14.8	24.0	24.6
SUM OF SQUARES DEVIATION	226.7	226.4	211.9	202.6	202.0
GOODNESS OF FIT	.000	.001	.065	.106	.108
COEF. OF MULTIPLE CORR.	.000	.033	.255	.325	.329



CUBIC POLYNOMIAL : $Z = a_0 + a_1 d + a_2 d^2 + a_3 d^3$

where Z = log (Th/U) and d = depth (standardized as true depth - 2650 / 1000)

From regression analysis : $a_0 = 0.184$, $a_1 = 3.228$, $a_2 = -31.469$, $a_3 = 59.371$

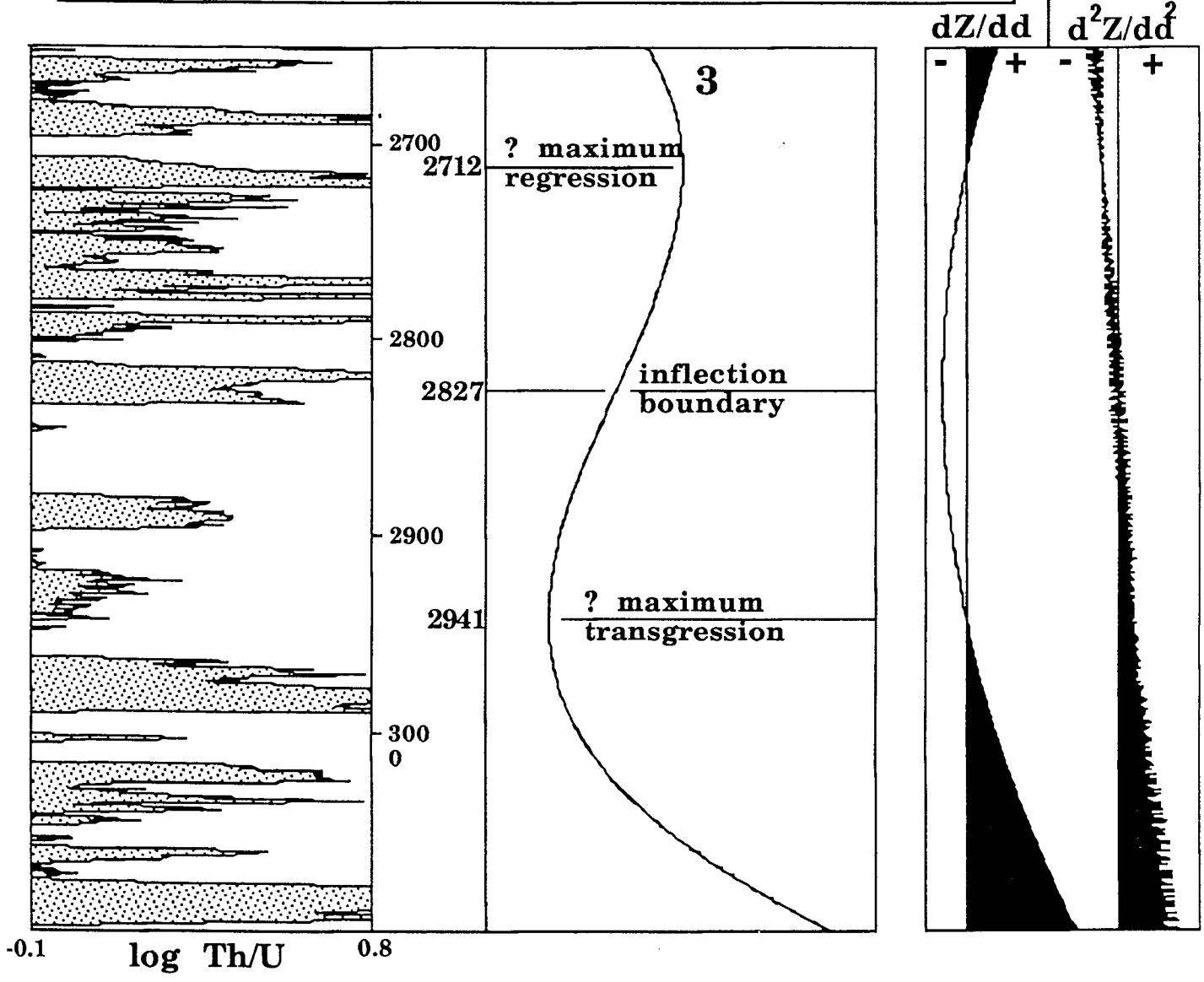
First derivative (slope) = $dZ/dd = a_1 + 2a_2 d + 3a_3 d^2$

$dZ/dd = 0$ at maxima or minima, and gives 2712 and 2941 feet

Second derivative = $d^2Z/dd^2 = 2a_2 + 6a_3 d$

$d^2Z/dd^2 = 0$ at inflection points, and gives one solution at 2827 feet.

OR "Quick - and - dirty"....
Use first difference of cubic trend curve for dZ/dd by applying a FILTER with elements (-1, 1)
and second difference for d^2Z/dd^2 using FILTER (1,-2,1)



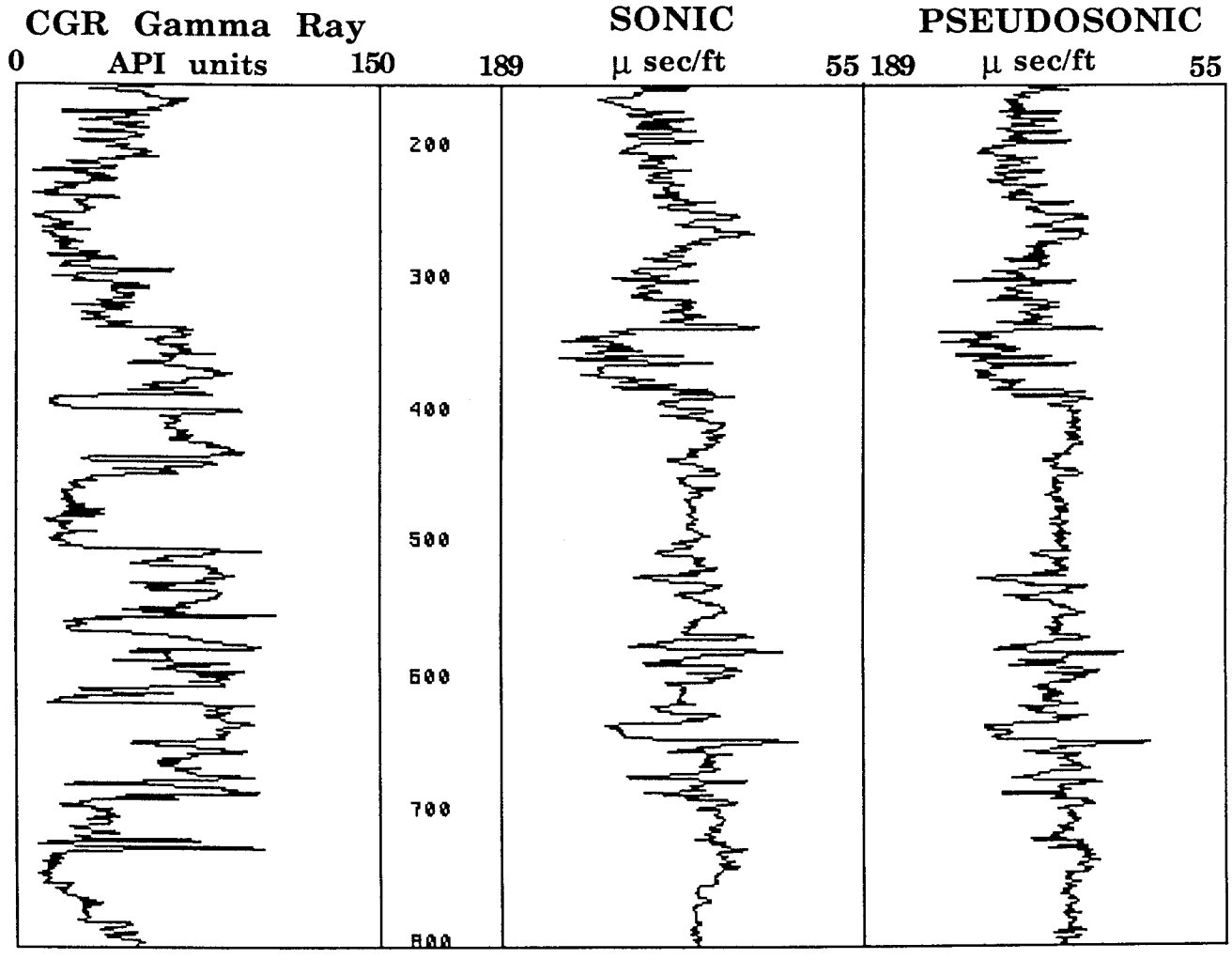
MULTIPLE REGRESSION : PREDICTION

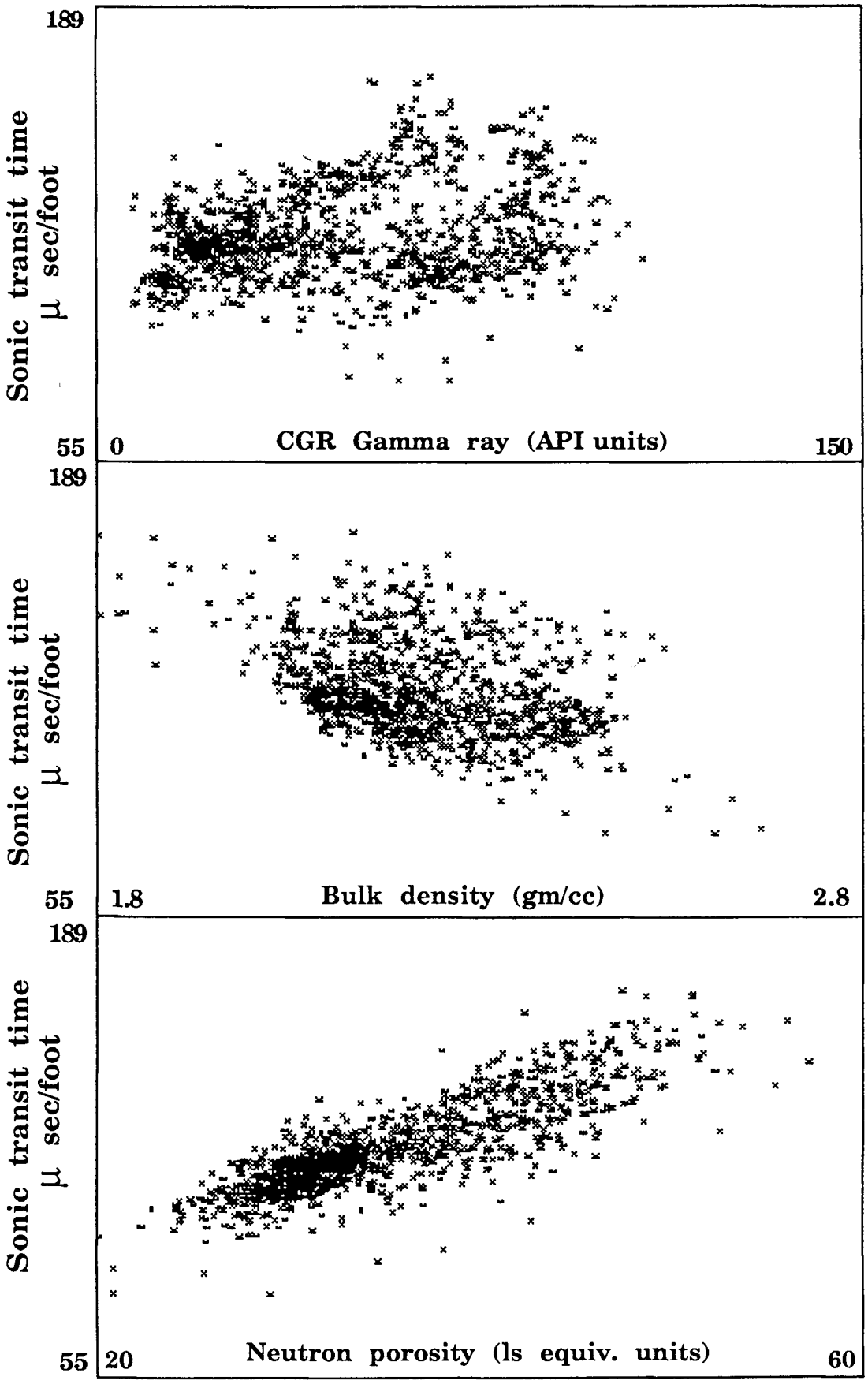
Development of prediction equation to generate pseudo-sonic logs based on regression of sonic transit time on gamma-ray, density, and neutron logs.

$$\hat{\Delta t} = a_0 + a_1 G + a_2 \rho_b + a_3 \Phi_n$$

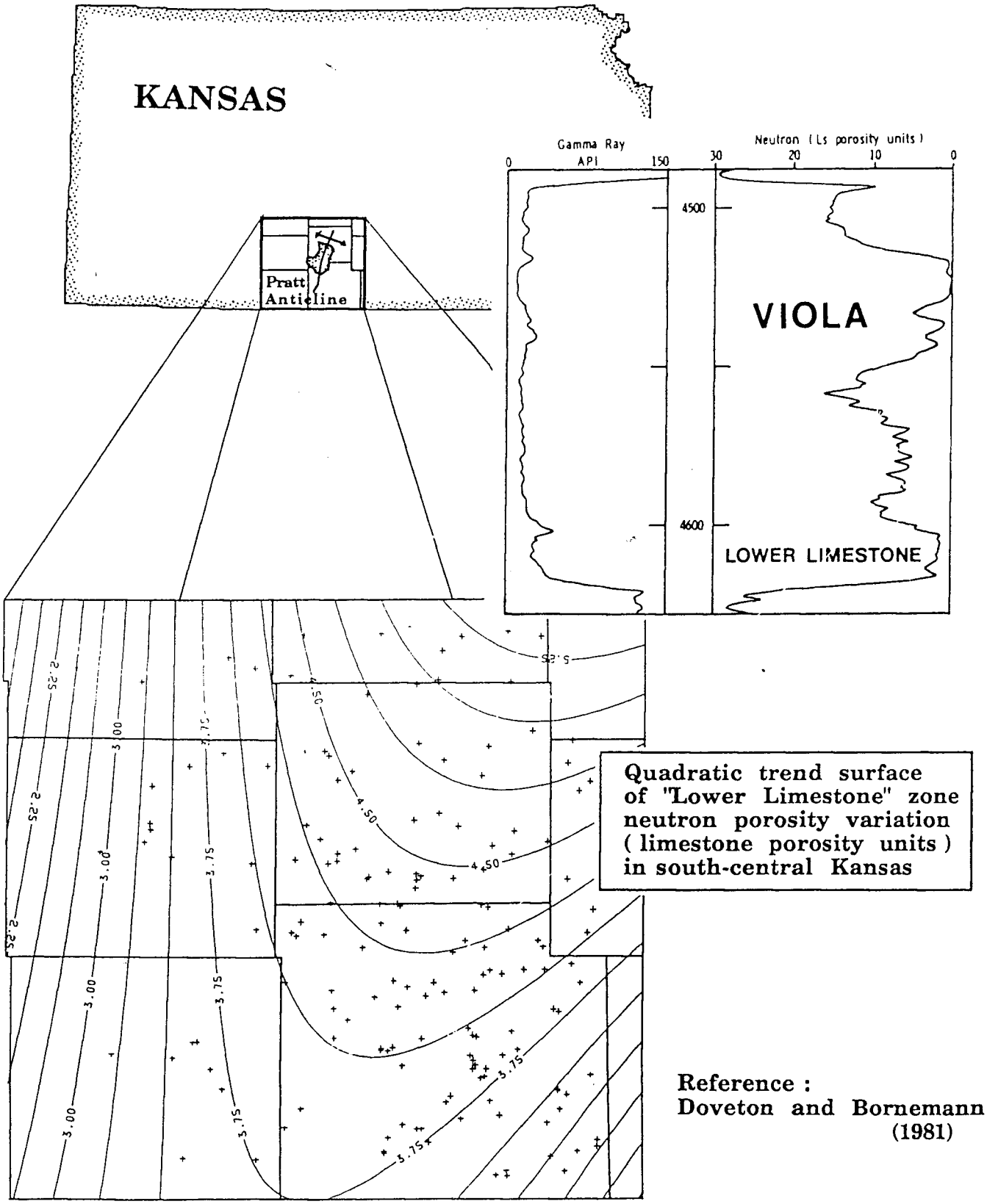
```
WELL NAME:
LOCATION:
DATE:
DEPTH:        150.00 TO    800.00 BY    .50 FEET
DEPENDENT VARIABLE:      DT

MULTIPLE REGRESSION

CONSTANT                  77.195        .000
CGR                         .029        .054
RHOB                      -7.425       -.068
NPHI                       1.689        .807

            SUM OF SQUARES    DEGREES-OF-FREEDOM    MEAN-SQUARES    F-TEST
REGRESSION     162579.81                     3        54193.269  1139.192
DEVIATION       61700.43                  1297           47.571
TOTAL          224280.25                  1300 ....
```

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



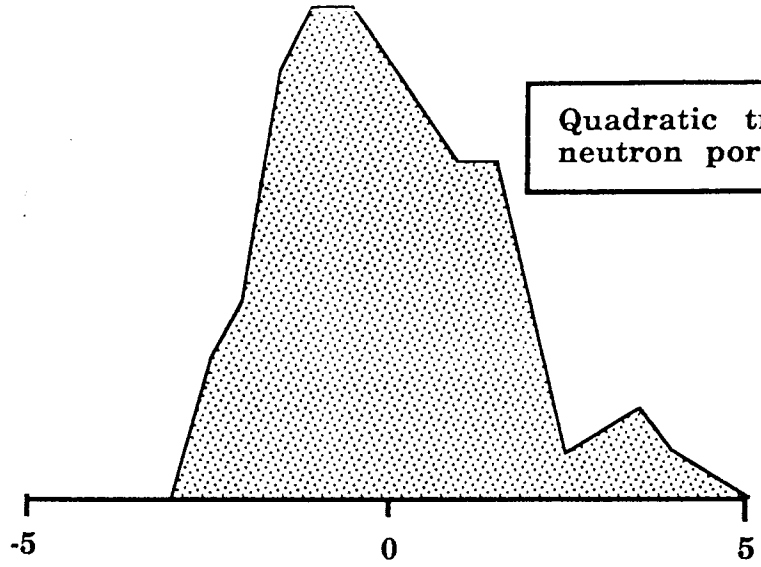

189
Sonic transit time
μ sec/foot
55
0
CGR Gamma ray (API units)
150
189
Sonic transit time
μ sec/foot
55
1.8
Bulk density (gm/cc)
2.8
189
Sonic transit time
μ sec/foot
55
20
Neutron porosity (ls equiv. units)
60

LOG NORMALIZATION BY TREND - SURFACE ANALYSIS



Quadratic trend surface of "Lower Limestone" zone neutron porosity variation (limestone porosity units) in south-central Kansas

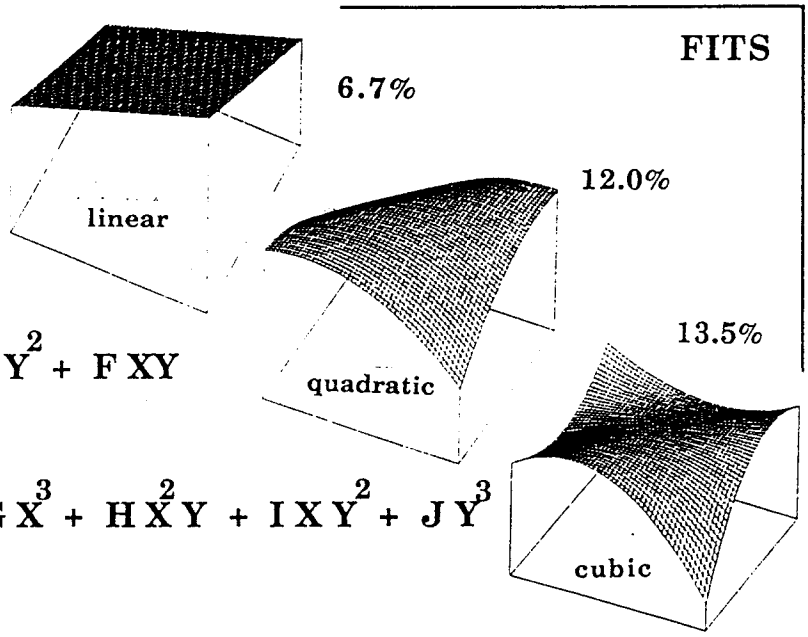
Reference :
Doveton and Bornemann (1981)



Quadratic trend-surface neutron porosity residuals

FITS

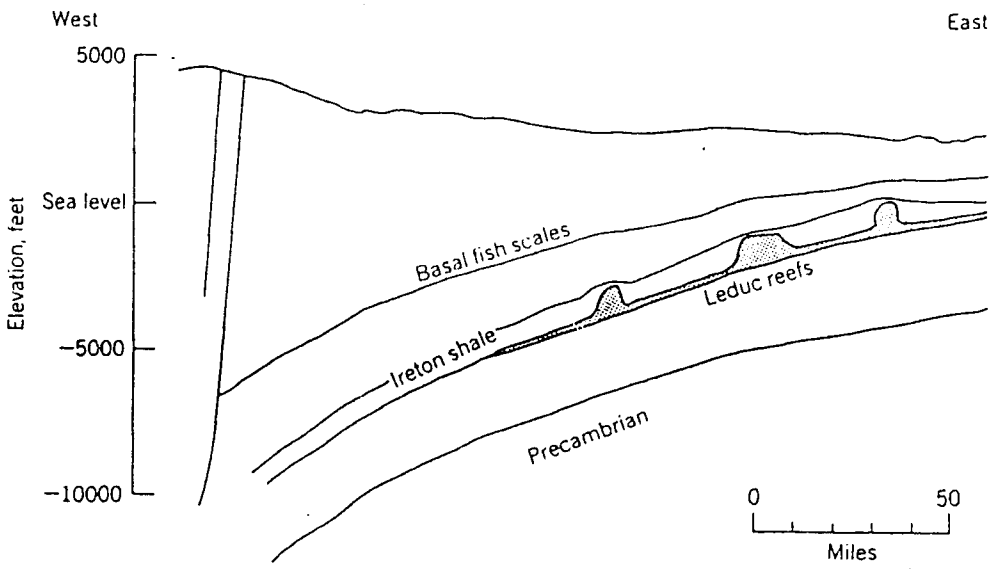
$$\hat{\Phi}_n = A + BX + CY + DX^2 + EY^2 + FXY + GX^3 + HX^2Y + IXY^2 + JY^3$$



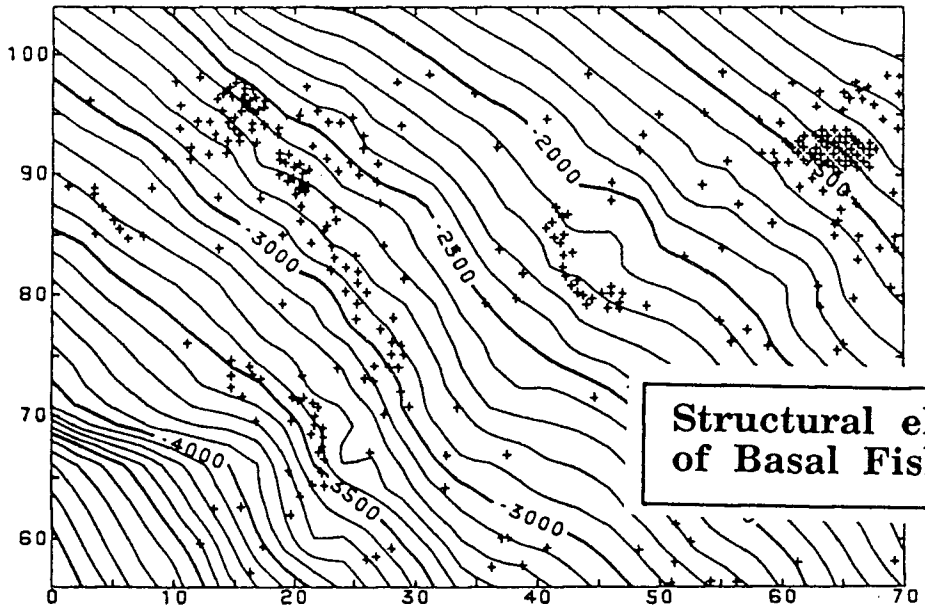
Source of Variation	Sum of Squares	DF	Mean Squares	F Ratio
Linear regression	27.22	2	13.61	5.73*
Linear deviation	377.62	159	2.37	
Quadratic–linear regression	21.41	3	7.14	3.13*
Quadratic deviation	356.22	156	2.28	
Cubic-quadratic regression	5.99	4	1.50	0.65
Cubic deviation	350.23	152	2.30	
Total variation	404.85	161		

*Significant at 5 percent level.

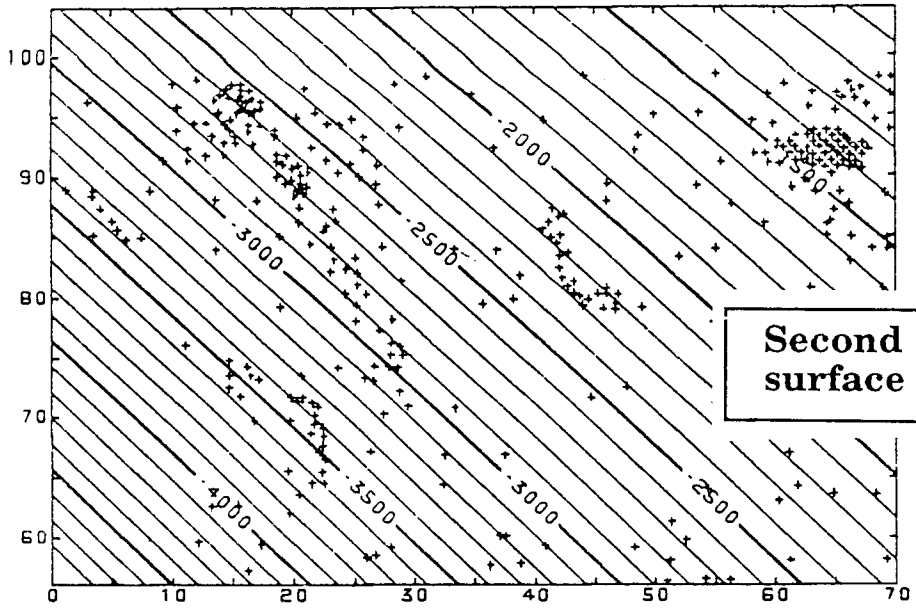
TREND SURFACE ANALYSIS OF STRUCTURE



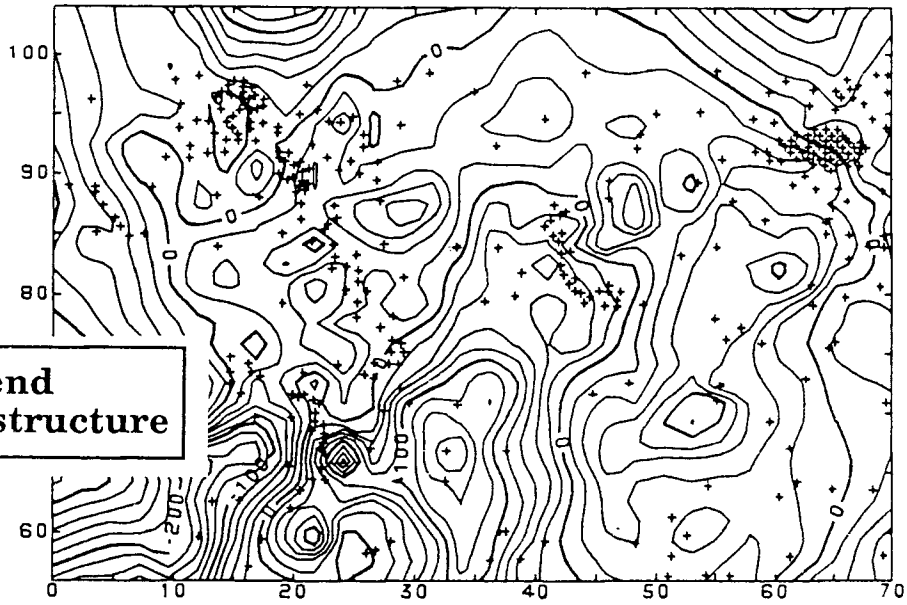
Cross-section of western Alberta, showing structural drape of Basal Fish Scales (Lower Cretaceous) over Leduc reefs (Upper Devonian)



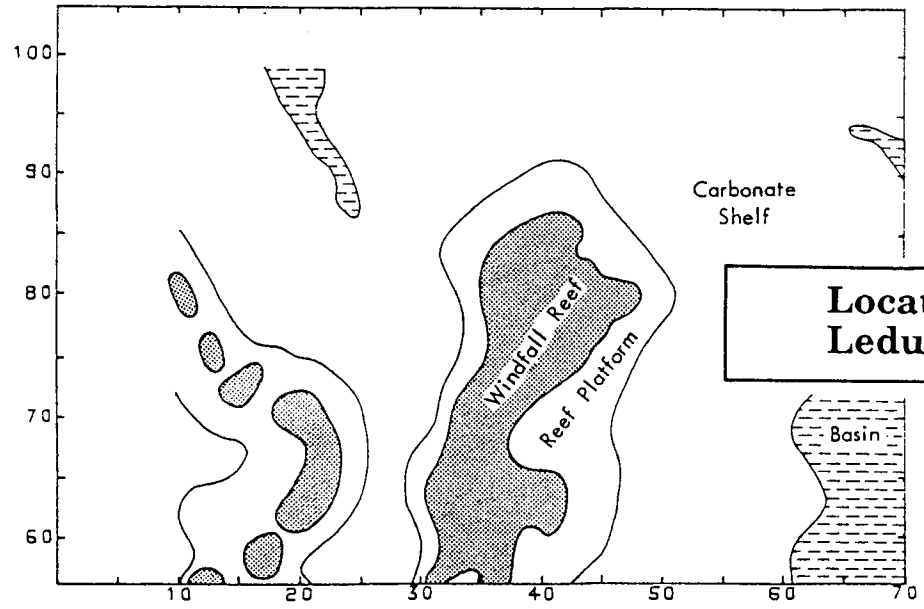
Structural elevation of Basal Fish Scales



Second degree trend surface of BFS structure

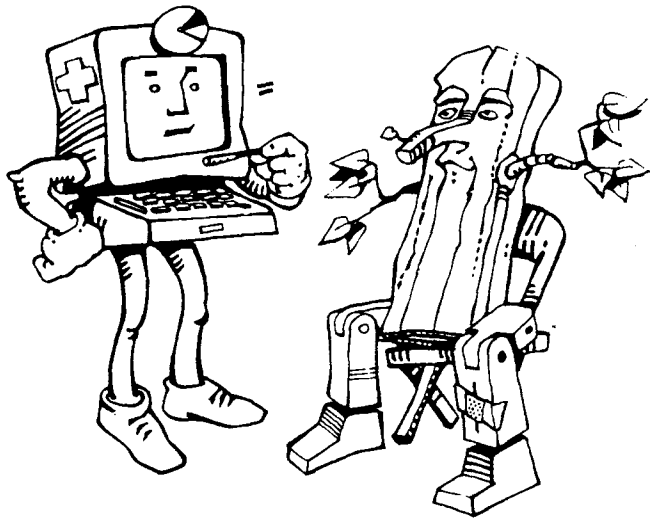


Second degree trend residuals of BFS structure



Location of known Leduc reefs

COMPUTER ANALYSIS
OF WELL LOG DATA



"Temperature normal... sap pressure normal... except for that nasty bark on your shin I'd say you're a well log."

(GEOBYTE)