Fall Semester 2005
University of Kansas
Department of Chemical and Petroleum Engineering

C&PE 940

Petroleum Geoscience Data Analysis

John H. Doveton

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.

The TORIS database contains variables on all four measurement scales. The reservoir properties in TORIS.xls are listed below, together with their scale of measurement:

State Postal Code	NOMINAL
Lithology Code	NOMINAL
Geologic Age Code, AAPG	ORDINAL
Net Pay (Feet)	RATIO
Gross Pay (Feet)	RATIO
Porosity (%)	RATIO
Initial Water Saturation (%)	RATIO
True Vertical Depth (Feet)	INTERVAL
Formation Temperature (degrees F)	INTERVAL
Current Formation Pressure (PSI)	RATIO
Permeability (MD)	RATIO
API Gravity (∫API)	INTERVAL
Formation Salinity (PPM TDS)	RATIO
OOIP (BBL)	RATIO
Reservoir Acreage (Acres)	RATIO
Initial Formation Pressure (PSI)	RATIO
Depositional System	NOMINAL

BASICS OF PROBABILITY

The earliest research and publications on probability were concerned with gambling, particularly in games that involved the results of throwing dice. The drilling of wildcat exploration wells in search of oil and gas is a gamble and the risk analysis used by oil companies is built on probability concepts to design exploration strategies and evaluate them in terms of economics.

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 (as in games of chance). *Empirical probabilities* are measured as a frequency ratio from an observed trial series where:

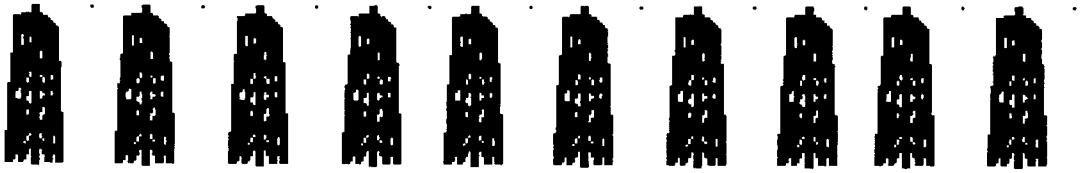
$$\text{probability of event} = \frac{\text{total number of occurrences of event}}{\text{total number of trials}}$$

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

Let us suppose that an oil company is drilling wildcat exploration holes in a new area where the probability of success is considered to be $P = 0.1$. In other words, on the average, only one in ten wildcats is expected to discover a new field. The company decides to launch a drilling program of ten wells. The company might expect to make one discovery, but would realize that all ten holes could be dry or there might be several discoveries. How can probability be used to quantify predictions for the drilling program?

If p is the probability that an event occurs ("success") and q is the probability that it does not occur ("failure"), then $p = 1 - q$. There are two alternative events for any trial and the system is **binomial**.

The probability of an event not occurring in n trials is q^n. So, the probability of no discoveries for the ten wildcat series, when $P = 0.1$ (and $q = 0.9$) is 0.349.

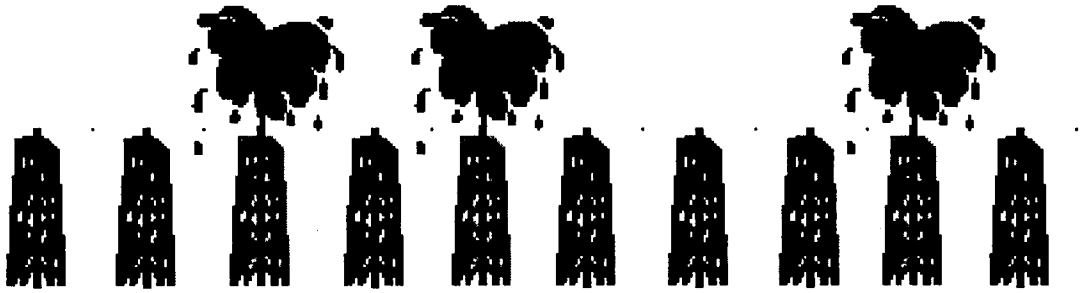


We can also ask the question of how many wildcats do we need to drill before we can be 90% certain that we will have at least one discovery. This is equivalent to saying what is the total number of wildcats to be drilled before the probability of all dry holes drops below 0.1 or $q^n < 0.1$. Since the critical value for n occurs when $n = \log p / \log q$ then $n = 21.85$. So, at least 22 wildcats would need to be drilled.

The probability of a "success" on the nth trial (but on none of the previous) is $q^{n-1}p$. However, the probability of the success occurring only once in n trials (regardless of which trial) is $nq^{n-1}p$ since there are n different positions for this success. So, the probability of the first wildcat resulting in a discovery is 0.039, while the probability of one discovery somewhere in the drilling program is 0.387. (Notice that the probability of at least one discovery is 0.651 because the probability of ten dry holes is 0.349).

The probability of there being n-r failures followed by r successes is $q^{n-r}p^{r}$. However, the n-r failures and r successes can be arranged in $\frac{n!}{(n-r)!r!}$ ways which are mutually exclusive. So, the probability of n-r failures and r successes (without regard to order is $\frac{n!}{(n-r)!r!}.q^{n-r}p^{r}$. This equation is the description of the binomial distribution. Substitution of n, p, and successive integer values of r into the equation gives probability values of combinations of r successes and n-r failures.

A binomial distribution can be set up in an EXCEL worksheet using this equation or by using BINOMDIST from the Function Wizard. The result for the ten wildcat drilling program is shown on the next page, but the spreadsheet is generalized for different success probabilities and numbers of trials.

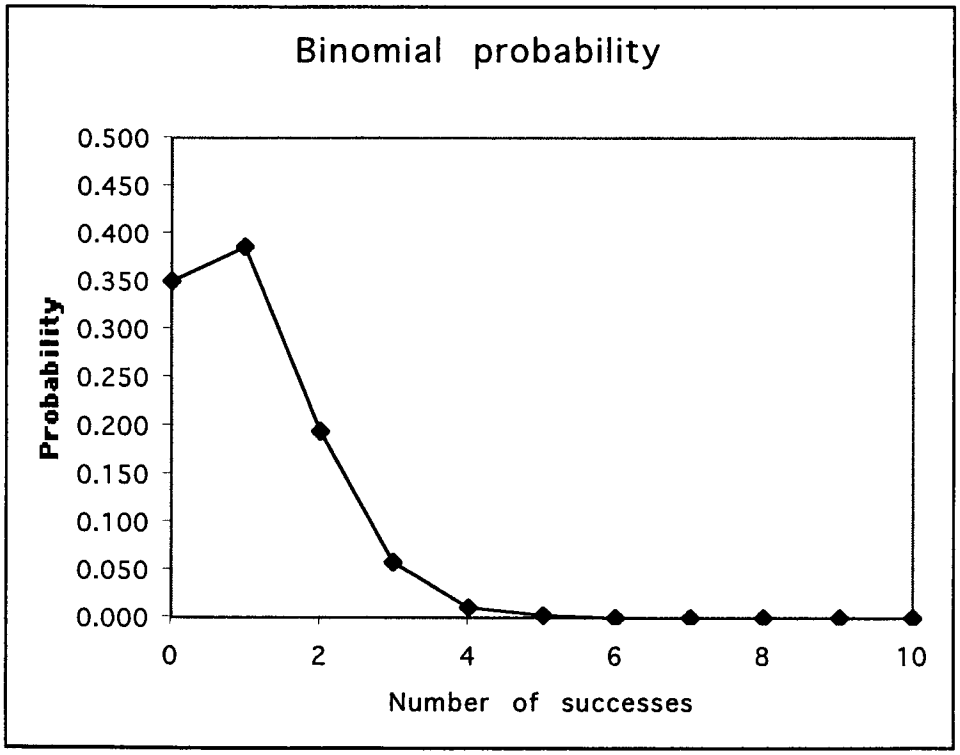


Binomial probability distribution

number of trials	10
probability of success	0.1

successes	probability
0	0.349
1	0.387
2	0.194
3	0.057
4	0.011
5	0.001
6	0.000
7	0.000
8	0.000
9	0.000
10	0.000

BINOMDIST(s,n,p,c)
s = number of successes
n = number of trials
p = probability of success
c = FALSE



EXCEL spreadsheet for binomial distribution using probability of number of discoveries in a wildcat drilling program of independent events.

In the real world, wildcats are not drilled randomly, but use information gained from geology and geophysics. However, the probabilities that are calculated for independent events are a useful baseline of worst-case scenario. The incorporation of geological information can improve these odds and also be included within an analysis of probabilities.

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*. In the wildcat drilling program, the events of success or failure were mutually exclusive.

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 of the northern U.S.. When wildcats are drilled, the potential reservoir formations penetrated are evaluated as to their likelihood of commercial oil or gas production. Hydrocarbon indications fro the drill-cutting or geophysical logs may be sufficiently encouraging to run a drill-stem that will produce fluids from the formation. If the test suggests that there are enough hydrocarbons to be produced economically, then the wildcat will be completed as an oil or gas well.

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. and the six outcomes were condensed to their categories of "commercial potential" (C) and "non-commercial" (N).

CONDITIONAL PROBABILITY ANALYSIS

DST results (C = commercial hydrocarbons; N = non-commercial)
versus SW (water saturation) estimated from logs
Data from Williston Basin carbonate reservoirs (Teti and Krug, 1987)

SW	C	N
0-9	0	0
10 to 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	0	6
80-89	0	2
90-100	0	4
	16	53

P(C)	P(N)
0.232	0.768

OBSERVED TALLY

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

OBSERVED PROBABILITY

	C	N
Sw<50%	0.203	0.333
Sw>50%	0.029	0.435

EXPECTED IF INDEPENDENT

	C	N
Sw<50%	0.124	0.412
Sw>50%	0.108	0.356

Bayesian prediction
when wildcat success p = 0.1

	P (SW/C)	P (SW/N)
SW<10	0.000	0.000
SW<20	0.125	0.019
SW<30	0.375	0.075
SW<40	0.563	0.245
SW<50	0.875	0.434
SW<60	0.938	0.566
SW<70	1.000	0.774
SW<80	1.000	0.887
SW<90	1.000	0.925
SW<100	1.000	1.000

P(C)	0.100

SW<10	#DIV/0!
SW<20	0.424
SW<30	0.356
SW<40	0.203
SW<50	0.183
SW<60	0.155
SW<70	0.126
SW<80	0.111
SW<90	0.107
SW<100	0.100

EXCEL summary spreadsheet of DST results from the Williston Basin and associated probability calculations.

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. A number of wildcats will have been drilled where no DSTs were run at all because the indications of oil and gas were so poor in drilling the hole.

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 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.

INDEPENDENT PROBABILITY EXPECTATIONS OF CATEGORY FREQUENCIES COMPARED WITH OBSERVED FREQUENCIES IN DATA TABLES

Comparisons can be made between observed frequencies of mutually exclusive categories and the frequencies that would be expected if there was no association between the categories (independence). Differences between observed and expected reveal patterns of association that show systematic relationships and, if interpreted correctly, the nature of the causal variables. Contingency table analysis is particularly important when dealing with variables measured on either the nominal or ordinal scales.

The TORIS database of U.S. oilfields contains variables measured on both discrete and continuous scales. The lithologies of the reservoirs are subdivided between sandstones, limestones, and dolomites. These are nominal data (discrete and no order). Because there are so few dolomites represented in the database, we will combine dolomites and limestones in a single category of 'carbonates'. Entries in the geological age category are tabulated as a three-number code. When subdivided between Tertiary (youngest), Mesozoic, and Paleozoic (oldest), the measurement scale is ordinal (discrete, but ordered). The areal size of each oil reservoir is reported in acres and so is measured on a ratio scale (continuous and with an absolute zero). In order make frequency comparisons with the other two discrete variables, we subdivide the areal size range of fields into categories. In this simple demonstration, we use two categories of 'small' (less than the median - or 50 percentile - of field sizes) and 'large' (greater than the median).

The spreadsheet of the resulting TORIS database field size summary shows a tabulation of categorical frequencies by lithology, age, and size, with conversion into their corresponding marginal probabilities. In the next table, comparisons can be made between the observed joint occurrences of lithology and age and their expected frequencies for independence calculated by the product of their marginal probabilities multiplied by the total number of fields. The pattern shows that Tertiary and Mesozoic age reservoirs are primarily sandstone, while carbonate reservoirs are mostly restricted to the Paleozoic. If the analysis is expanded to include reservoir areal size, then the frequencies show a clear tendency for small sandstone reservoirs in the Tertiary and Mesozoic as contrasted with large carbonate reservoirs in the Paleozoic.

TORIS DATABASE OF U.S. OILFIELDS
FIELD SIZE

TALLIES

LITHOLOGY		AGE		SIZE	
sandstone	759	Tertiary	386	small	532
carbonate	303	Mesozoic	180	large	530
		Paleozoic	496		
				TOTAL	1062

MARGINAL PROBABILITIES

sandstone	0.715	Tertiary	0.363	small	0.501
carbonate	0.285	Mesozoic	180.000	large	0.499
		Paleozoic	496.000		

		OBSERVED	EXPECTED
sandstone	Tertiary	368	275.9
sandstone	Mesozoic	161	128.6
sandstone	Paleozoic	230	354.5
carbonate	Tertiary	18	110.1
carbonate	Mesozoic	19	51.4
carbonate	Paleozoic	266	141.5
	total	1062	

			OBSERVED	EXPECTED
sandstone	Tertiary	small	264	138.2
sandstone	Tertiary	large	104	137.7
sandstone	Mesozoic	small	76	64.4
sandstone	Mesozoic	large	85	64.2
sandstone	Paleozoic	small	98	177.6
sandstone	Paleozoic	large	132	176.9
carbonate	Tertiary	small	9	55.2
carbonate	Tertiary	large	9	55.0
carbonate	Mesozoic	small	7	25.7
carbonate	Mesozoic	large	12	25.6
carbonate	Paleozoic	small	78	70.9
carbonate	Paleozoic	large	188	70.6
		total	1062	

KEY to categories
Reservoir **LITHOLOGY** : either sandstone or carbonate (limestone + dolomite)
Geological **AGE**: Tertiary (Eocene, Oligocene, Miocene, Pliocene), Mesozoic (Permian, Triassic Jurassic, Cretaceous), Paleozoic (Cambrian, Ordovician, Silurian, Devonian, Carboniferous)
Field SIZE: small (less than 2240 acres), large (greater than 2240 acres), median = 2240 acres

In a second example from the TORIS database, the API gravities of the oils are related to reservoir lithology and age. The tabulation shown in the spreadsheet is set p in the same style as the previous example. Notice that the total number of fields in this dataset (985) is different from the lithology - age - size data set (1062) because of missing entries in some of the data fields.

The API gravity of an oil is an arbitrary and direct function of the oil density given by the equation:

API Gravity = (141.5 / SG at 60* F) - 131.5

As can be seen, an oil with a density of water (SG = 1) would have an API Gravity of 10°,. Oils heavier than water would be less than 10°; those lighter than water are greater than 10°,. As a general rule, higher API gravity degree oil values have a greater commercial value and lower degree values have lower commercial value in terms of refinery costs. When world oil prices are quoted, they use oil from the North Sea Brent Field as a standard ("Brent Crude") which has an API Gravity of 39.6°. By contrast, a bituminous oil such as the Athabasca tar sands of Canada does not flow at normal temperatures and has an API gravity of about 8°. The API gravity of an oil not only holds economic implications regarding its value to a refinery, but is also key input to reservoir simulations of recovery in primary, secondary, and tertiary operations.

A widely accepted subdivision of crude oils defines Light crude oil as having an API gravity higher than 31.1°, Medium oil as having an API gravity between 22.3° and 31.1°, and Heavy oil as having an API gravity below 22.3°. The API gravity values in the TORIS database were allocated between these three categories. When cross-tabulated with lithology, there is a preferential association of lighter oils with sandstone reservoirs and heavy oils with carbonates. When also considered in terms of age, the younger sandstone reservoirs have lighter oils as contrasted with heavy oils in the older carbonate reservoirs. The results seem reasonable in the sense that older oils will tend to be more mature and also that older rocks may tend to have deeper burial depths and so elevated temperatures and pressures to enhance thermal maturation and lower density oils.

The interpretation of these two probability analyses is speculative because of the introductory nature of these examples. Later in the manual we will examine the application of statistical tests (principally the chi-square test) in contingency table analysis to establish the significance (or lack thereof) in differences between observed frequencies and expectations from an independence model. Also, additional variables and other methods will be used in the search for systematic associations and their causes.

TORIS DATABASE OF U.S. OILFIELDS
API GRAVITY

TALLIES

LITHOLOGY		AGE		GRAVITY	
sandstone	718	Tertiary	379	heavy	158
carbonate	267	Mesozoic	169	medium	200
		Paleozoic	437	light	627
				TOTAL	985

MARGINAL PROBABILITIES

sandstone	0.729	Tertiary	0.385	heavy	0.160
carbonate	0.271	Mesozoic	0.172	medium	0.203
		Paleozoic	0.444	light	0.637

		OBSERVED	EXPECTED
sandstone	heavy	147	115.2
sandstone	medium	165	145.8
sandstone	light	406	457.0
carbonate	heavy	11	42.8
carbonate	medium	35	54.2
carbonate	light	221	170.0
	total	985	

			OBSERVED	EXPECTED
sandstone	Tertiary	heavy	115	44.3
sandstone	Tertiary	medium	103	56.1
sandstone	Tertiary	light	146	175.9
sandstone	Mesozoic	heavy	23	19.8
sandstone	Mesozoic	medium	27	25.0
sandstone	Mesozoic	light	104	78.4
sandstone	Paleozoic	heavy	9	51.1
sandstone	Paleozoic	medium	35	64.7
sandstone	Paleozoic	light	156	202.8
carbonate	Tertiary	heavy	5	16.5
carbonate	Tertiary	medium	4	20.9
carbonate	Tertiary	light	6	65.4
carbonate	Mesozoic	heavy	1	7.3
carbonate	Mesozoic	medium	2	9.3
carbonate	Mesozoic	light	12	29.2
carbonate	Paleozoic	heavy	5	19.0
carbonate	Paleozoic	medium	29	24.1
carbonate	Paleozoic	light	203	75.4
		total	985	

KEY to categories
Reservoir **LITHOLOGY**: either sandstone or carbonate (limestone + dolomite)
Geological **AGE**: Tertiary (Eocene, Oligocene, Miocene, Pliocene), Mesozoic (Permian, Triassic Jurassic, Cretaceous), Paleozoic (Cambrian, Ordovician, Silurian, Devonian, Carboniferous)
API **GRAVITY**: heavy (less than 22.3), medium (between 22.3 and 31.1), light (greater than 31.1)

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*.

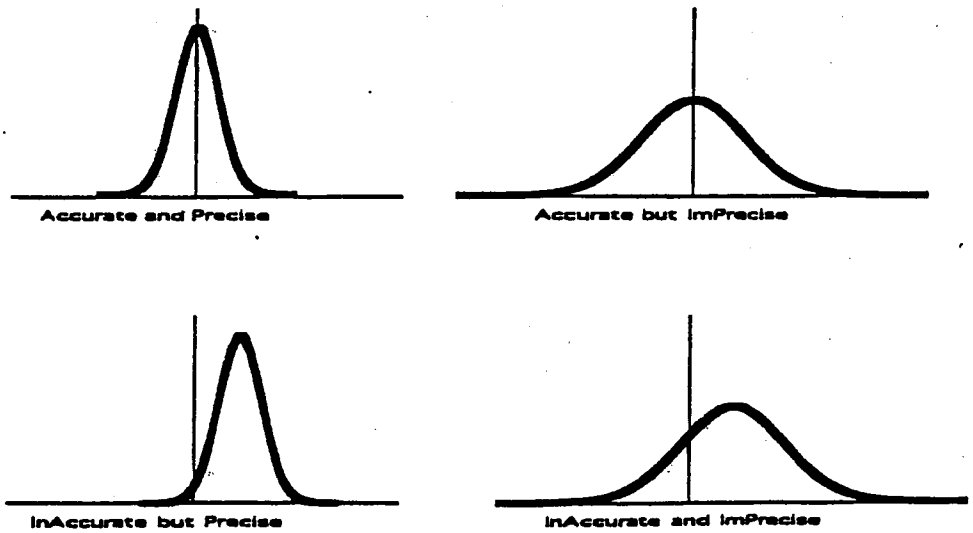
Univariate statistics are concerned with summarization and inferential analysis of a single variable measured on a sample of objects. *Multivariate statistics* marks an extension to several variables of measurement on each object and is the numerical description of variable interrelationships and inferences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

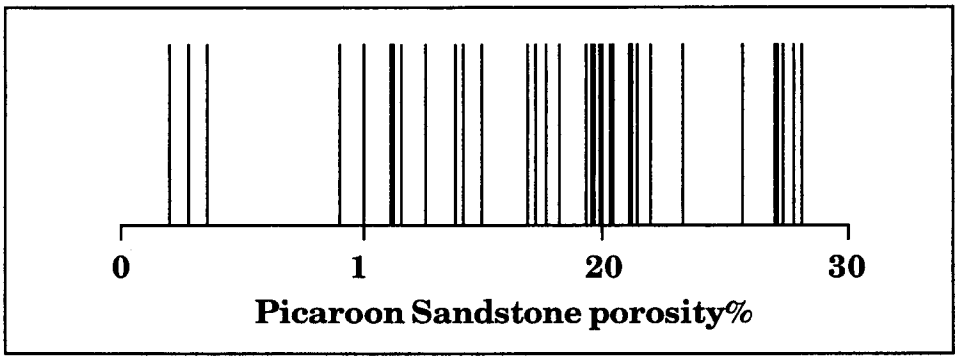
The terms *"accuracy"* and *"precision"* are sometimes (and wrongly) used interchangeably, but they have radically different meanings. The picture below should be helpful to clarify the distinction. Accuracy is a measure of how close to the true value the measured estimates tend to be. Precision is a 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.



GRAPHICAL DISPLAYS OF UNIVARIATE DATA

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



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.

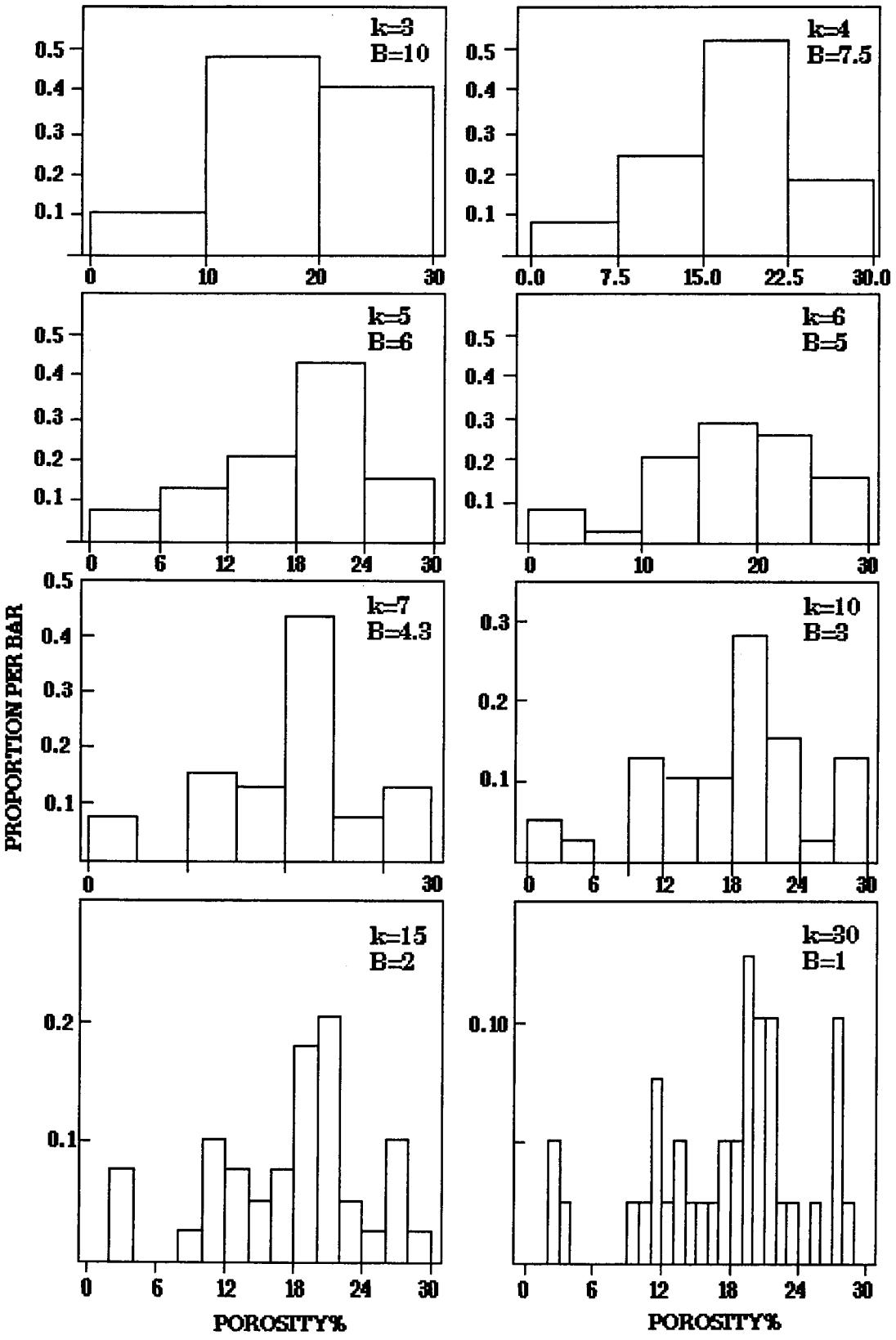
Too coarse a bin width causes an oversmoothing of the data; too narrow a bin results in poor generalization of the data density. This can be seen from a comparison of alternative histograms drawn for the Picaroon sandstone porosities. A number of rules have been devised to select appropriate bin width.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

$$k = \log_2 n$$

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.

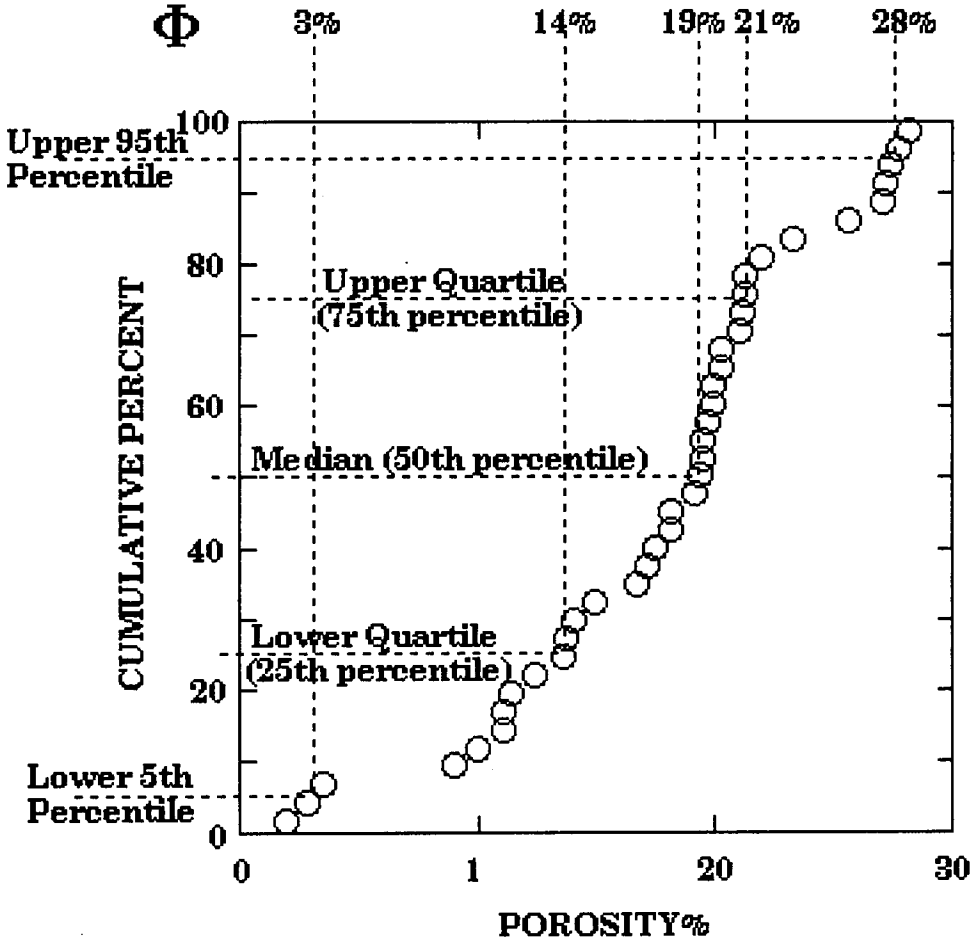
The edges of the bins mark abrupt discontinuities in the subdivision frequency counts of what is a continuous variable. Changes in boundary locations may adversely affect the histogram display, so that features may be less real than they appear. Some of these artifacts can be seen in the comparison of 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.



Alternative Picaroon Sandstone porosity histograms set by number of bins (k) and matching bin-widths (B)

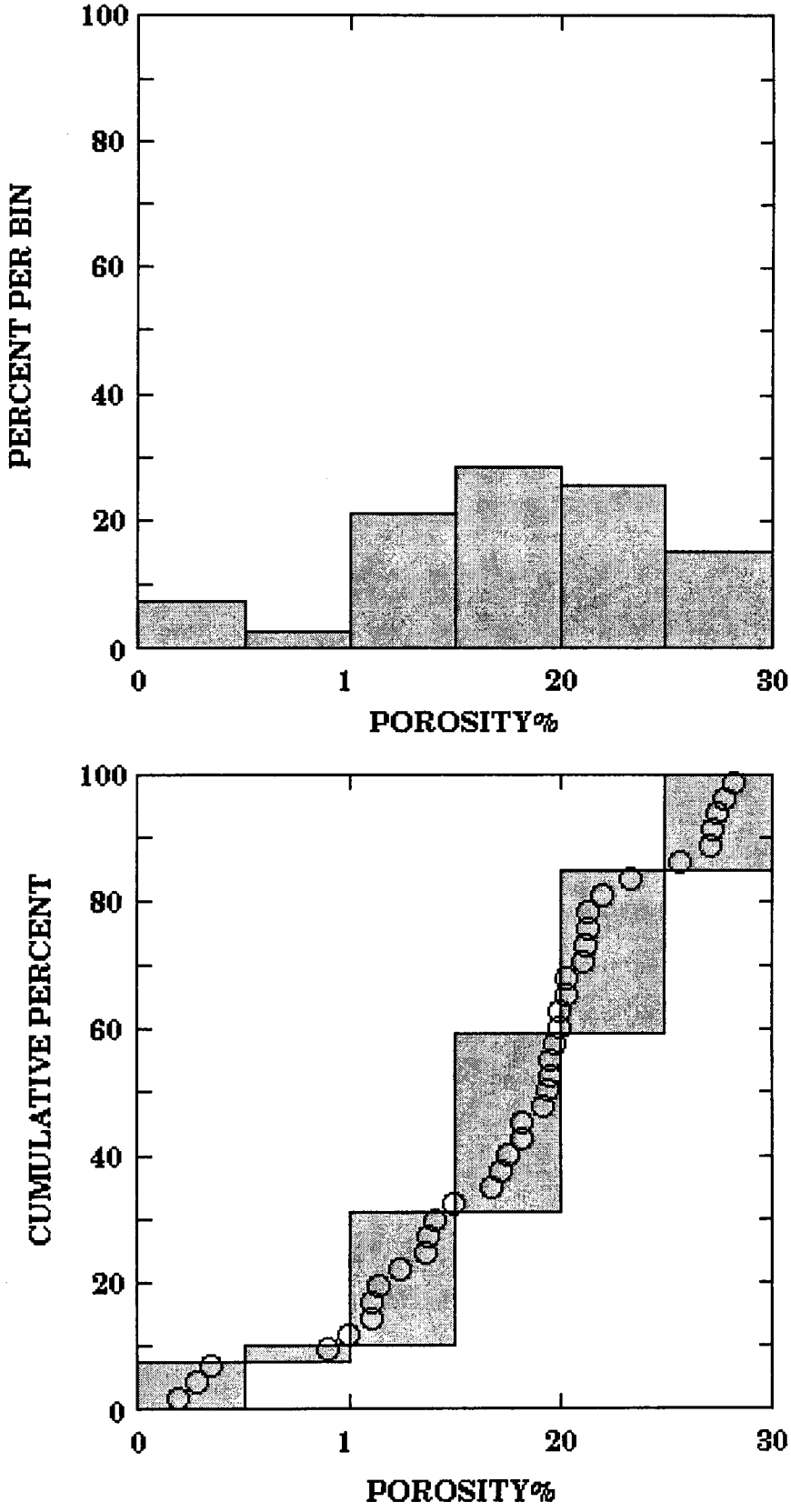
A quantile plot (abbreviated as Q-plot and also known as a cumulative plot) shows the individual observations plotted against their cumulative occurrence. A "quantile" of a sample is the value of the observation that matches a given proportion of the sample. A Q-plot of the Picaroon sandstone porosities 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; 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).

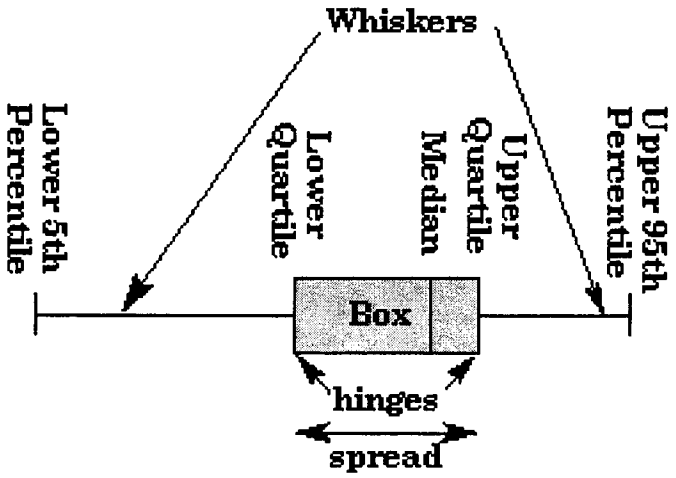
The Q-plot Picaroon sandstone porosities show a generalized ogive shape that contrasts the bulk of the central observations with the less frequent observations in the tails. Note the extended and particularly steep trend at about 20% porosity that marks the high proportion of observations close to this 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.



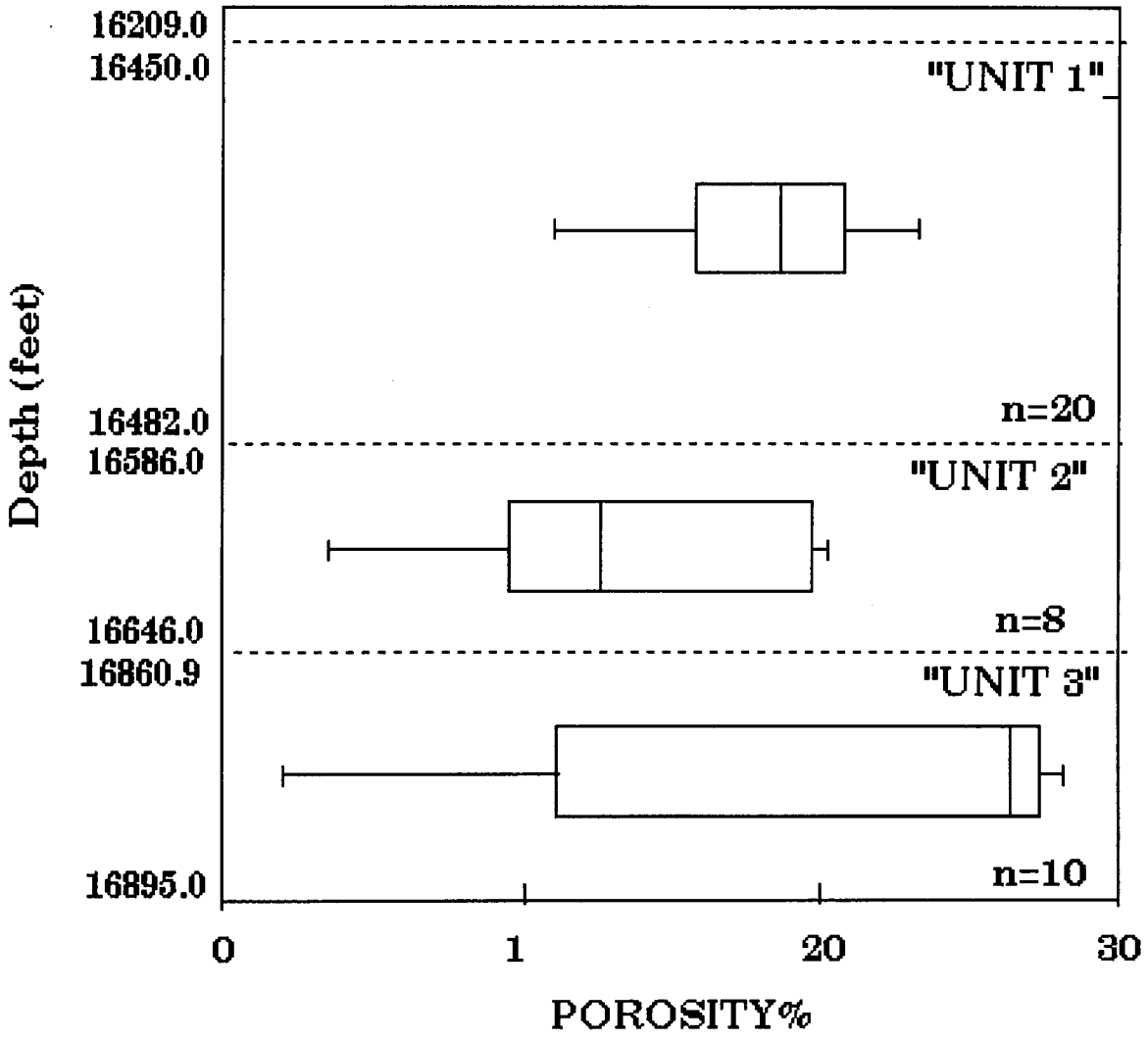
Relationship between histogram and cumulative plot for Picaroon Sandstone porosities

BOX PLOTS

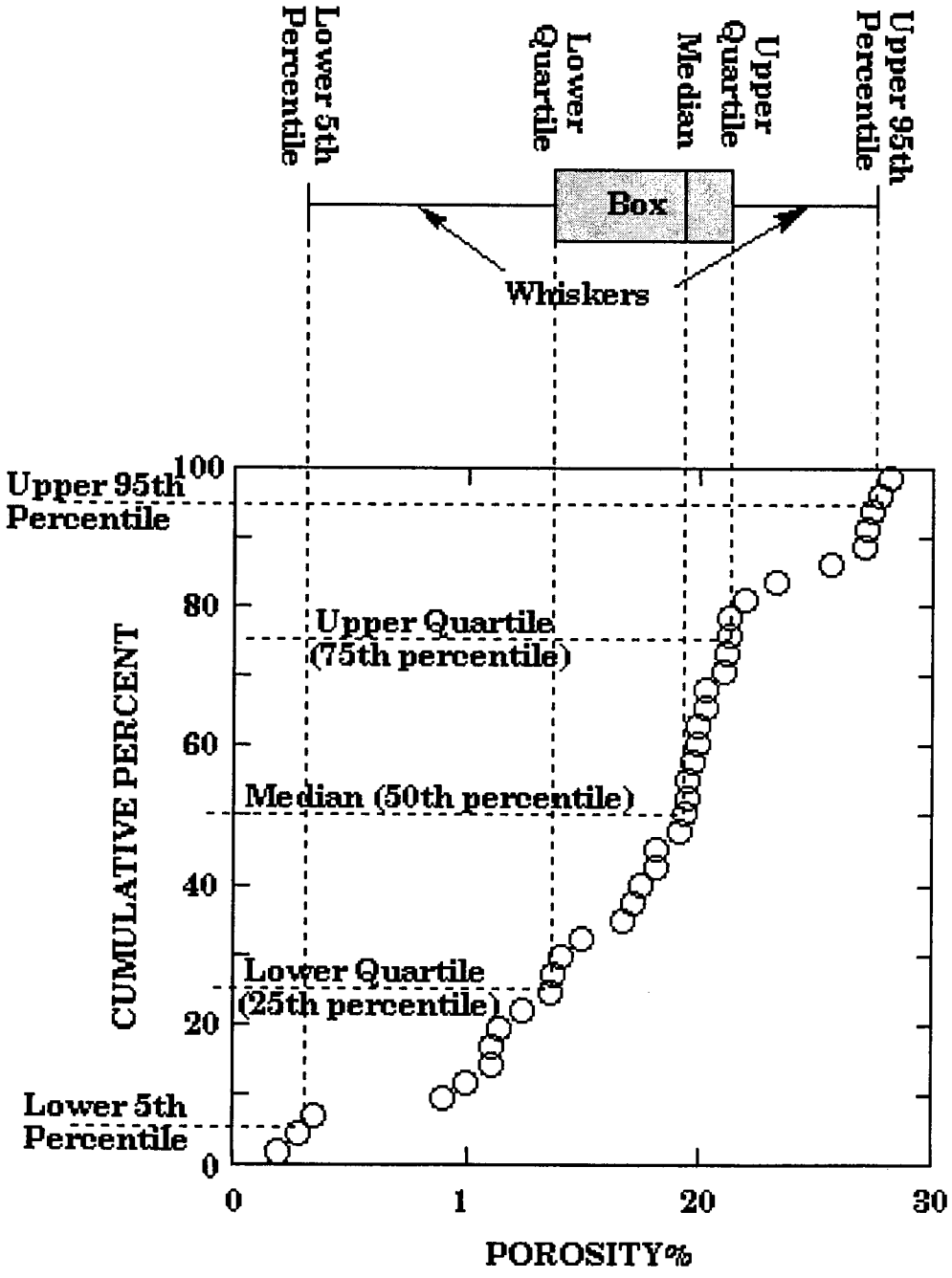
The method of box plots were introduced by Tukey (1977) as a simple graphical means to show the distribution form of a single variable. There are a variety of conventions on how box plots are drawn, but the one used to illustrate Picaroon sandstone porosity is fairly typical. The box features are drawn 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 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 (16450 to 16482 feet depth) show a compact, fairly symmetrical distribution centered on a porosity of about 18%. The boxes of "Unit 2" (16586 to 16646 feet depth) and "Unit 3" (16860.9 to 16895 feet depth) 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. 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.



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.

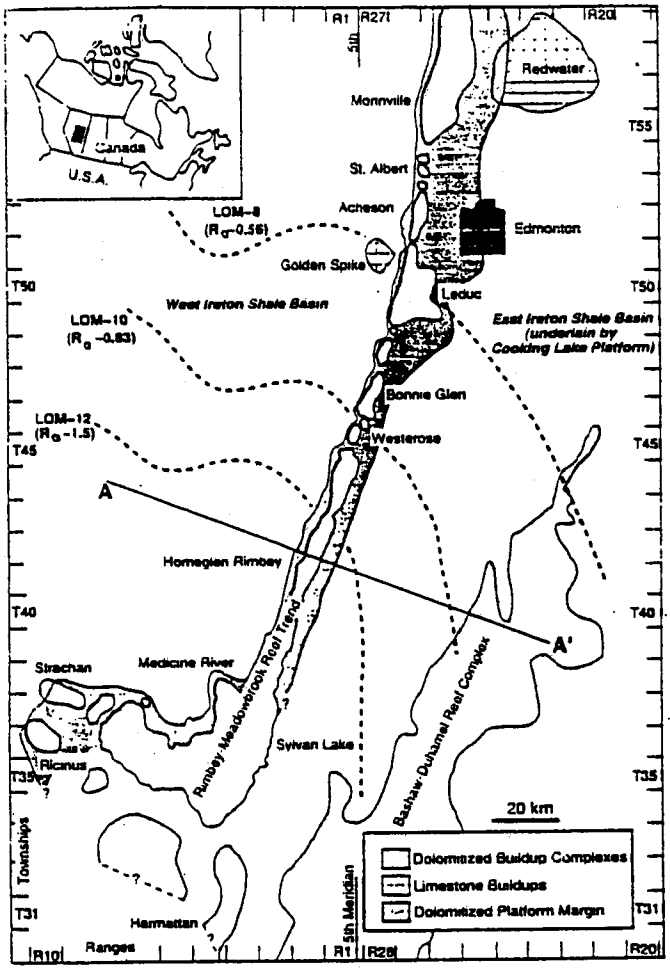
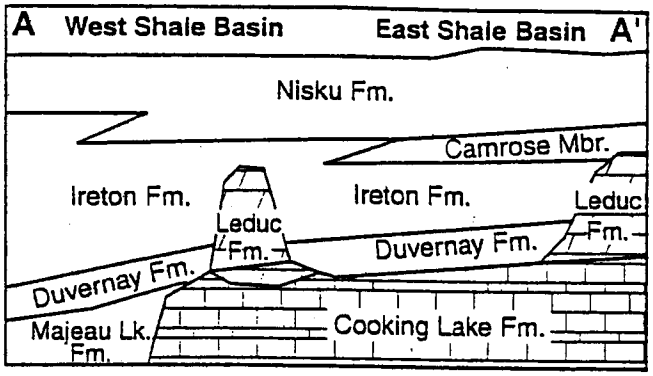
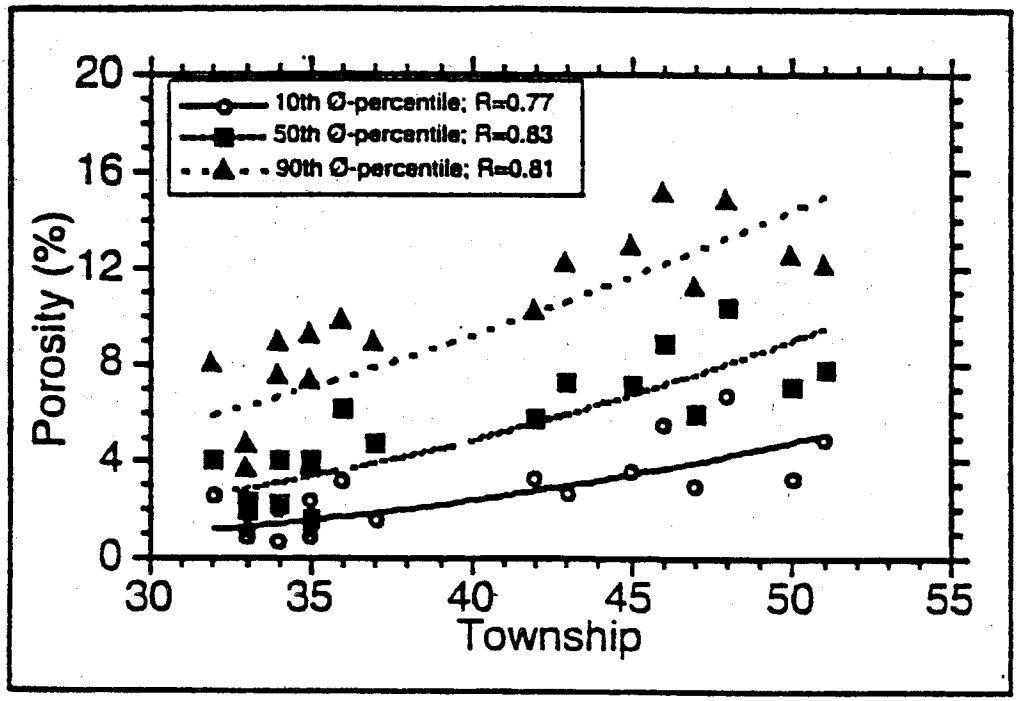


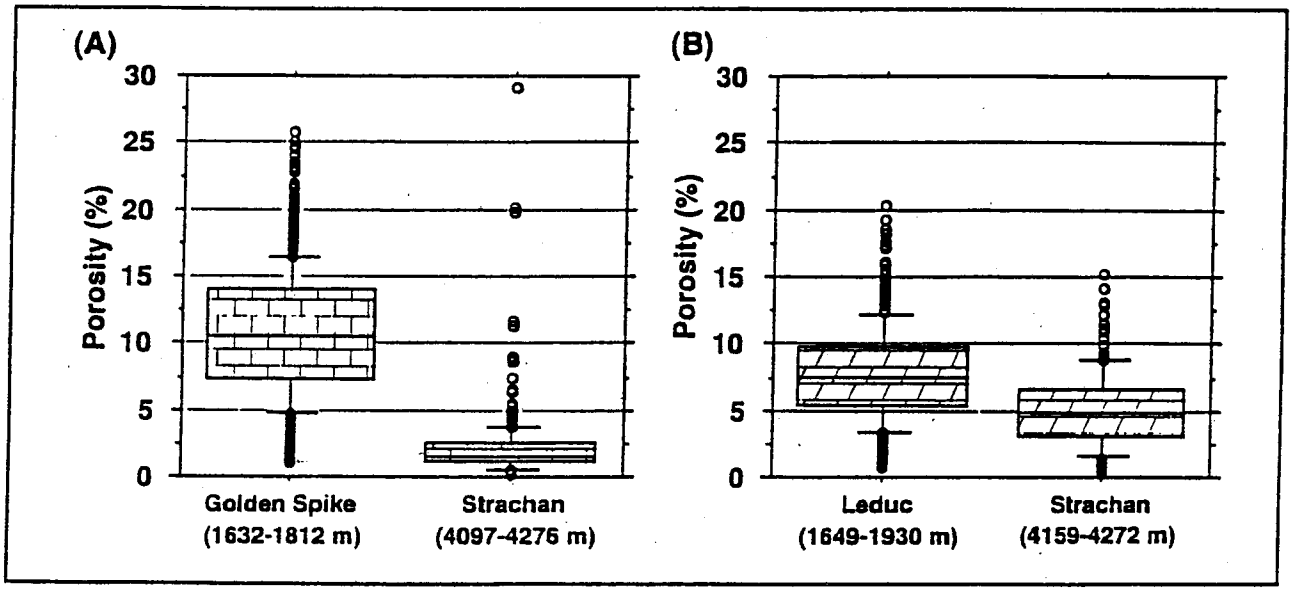
Figure 1—Map of the Rimbey-Meadowbrook reef trend, showing buildups and extent of dolomitization in the Cooking Lake carbonate platform (dark stippled pattern). Buildups are dolomitized where the margin of the Cooking Lake platform is dolomitized. Golden Spike and Redwater are situated off the margin 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).

from Amthor et al (1994)

BOX PLOTS IN EXCEL

Histograms and cumulative plots can be made in EXCEL using the Histogram option of the Data Analysis package listed under tools. Quartiles (or any percentiles) and the median are found by using functions listed under Function Wizard. An option for box plots is not available in EXCEL, but can be created easily using the simple spreadsheet Boxplot.xls downloaded from:

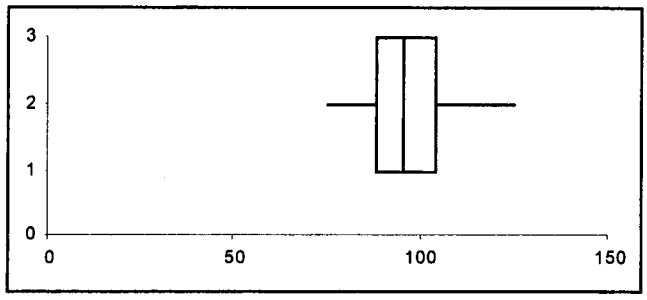
www.deakin.edu.au/~rodneyc/Boxplot.xls

The top of the spreadsheet looks like this:

Data
88
86
90
120
112
92
88
92
94
84
104
102
82
100
94
86
86
78
82
114
108
96
102
92
86

Min	75
Q1	88
Median	95
Q3	104
Max	125

x	y
75	2
88	2
88	3
95	3
95	1
88	1
88	3
104	3
104	1
88	1
88	3
104	3
104	2
125	2



The spreadsheet can be used as a template. Data are pasted into the Data column. The spreadsheet tabulates the minimum, lower quartile, median, upper quartile, and maximum values of the distribution. These numbers are used to create a box plot. The scale and size of the graphic can be altered using normal EXCEL tools. Notice that this box plot presentation uses the extreme values of the distribution as endpoints of the whiskers, rather than the 5% and 95% bounds. However, this can be changed easily by substituting percentile functions in the Min and Max cells.

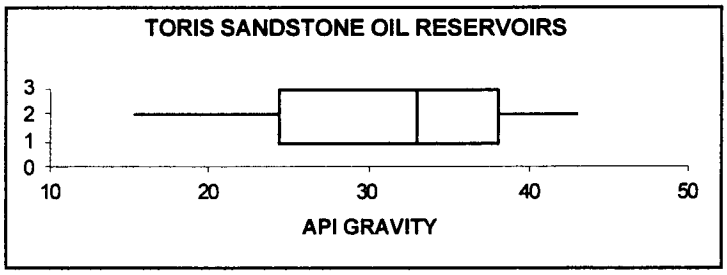
The spreadsheet can also be adapted to handle two (or more) data series as shown overleaf, where the distributions of API gravities of oils from sandstone and carbonate reservoirs are compared by box plots.

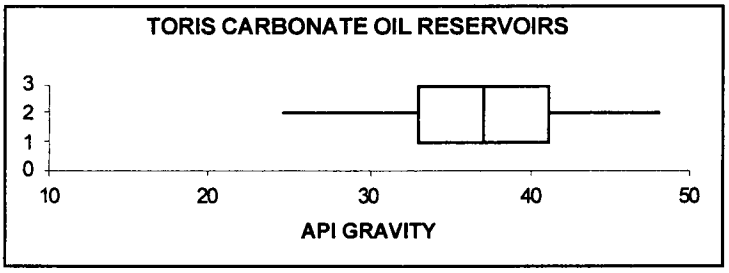
BOX PLOT
API GRAVITY OF TORIS RESERVOIR OILS

SS	CARB
6	4
7.5	13.6
8.5	14.9
9	18.6
9.5	19.3
10	20.1
10	20.5
10.6	20.9
12	21
12	21.5
12.3	21.5
12.5	22.5
12.7	22.7
13	23
13	23.7
13	24
13	24
13	24
13	25
13	25
13.3	25
13.5	26
13.5	26
13.5	26
14	27
14	27
14	27
14.2	27
14.2	27
14.3	27.9
14.4	28
14.5	28
15	28
15	28
15	28
15	28.2
15	28.6
15	28.8
15	29
15.1	29

	SS		CARB
5%	15.165	5%	24.5
Q1	24.325	Q1	33
Median	33	Median	37
Q3	38	Q3	41
95%	43	95%	48

x	y	x	y
15.165	2	24.5	2
24.325	2	33	2
24.325	3	33	3
33	3	37	3
33	1	37	1
24.325	1	33	1
24.325	3	33	3
38	3	41	3
38	1	41	1
24.325	1	33	1
24.325	3	33	3
38	3	41	3
38	2	41	2
43	2	48	2





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

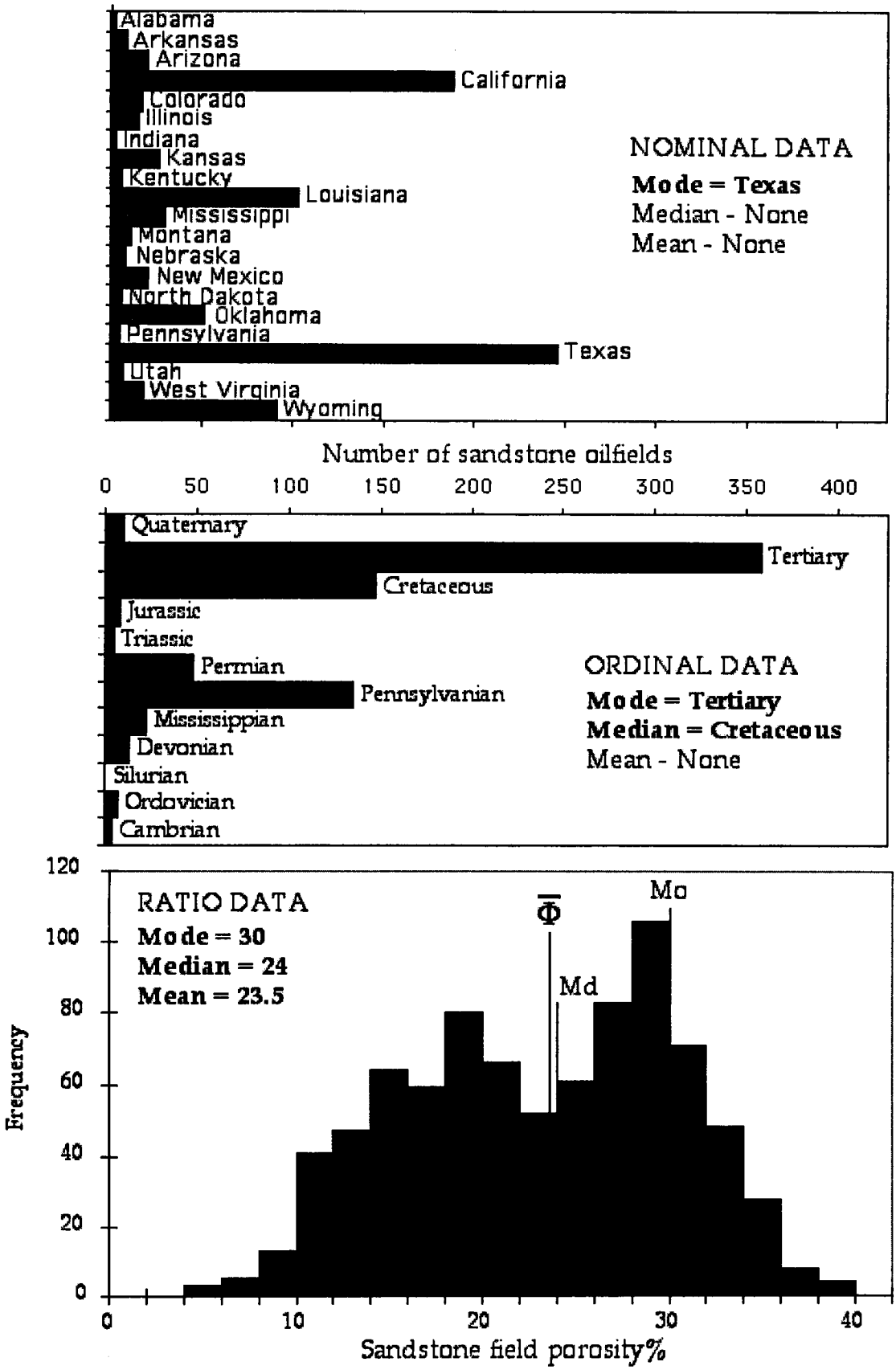
The mean of a sample is the arithmetic average, calculated by:

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

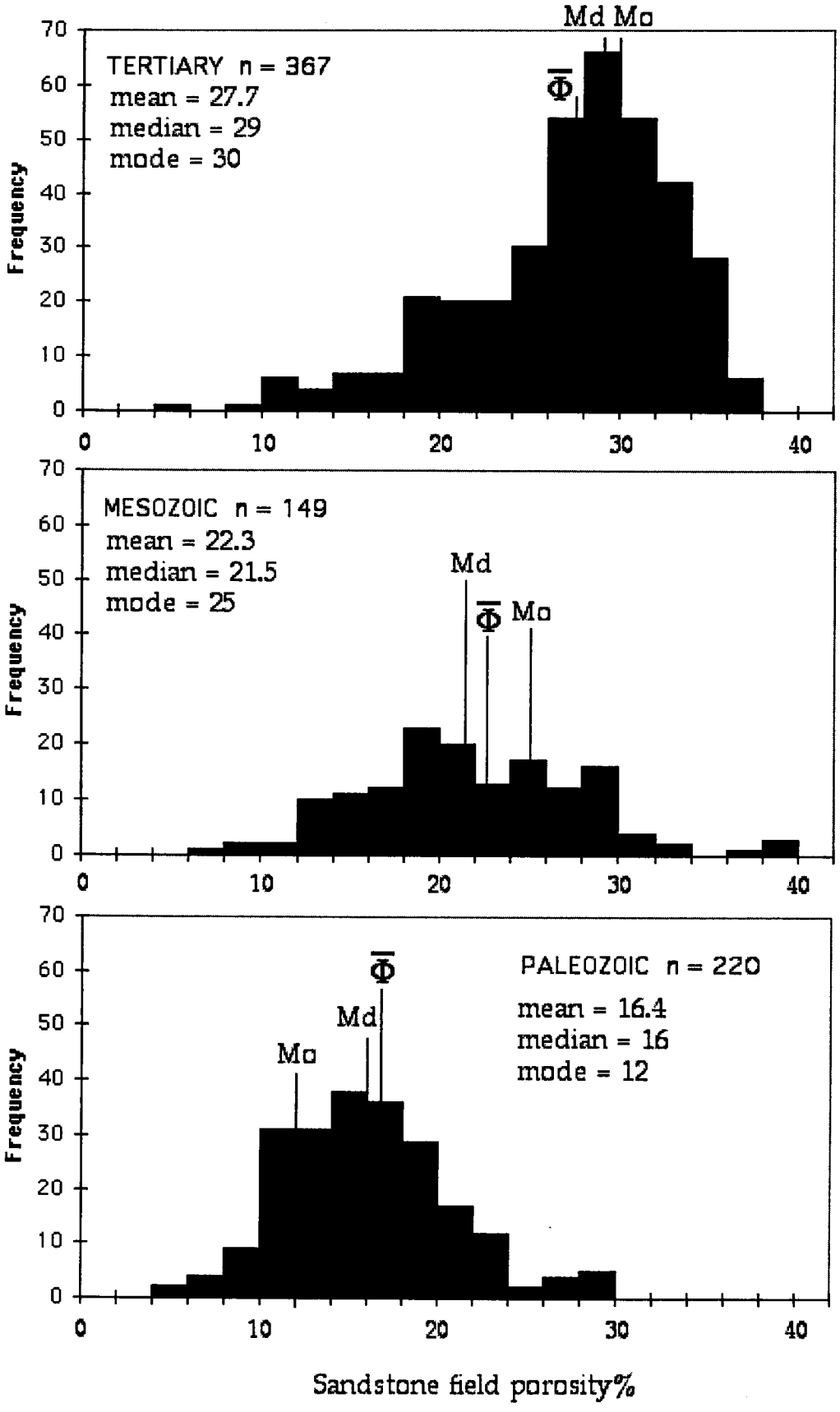
where n is the number of observations. The sample mean is an estimate of the population mean, μ. (Remember that sample estimates are in Latin script, population parameters are written in Greek.) In mechanical terms, the mean is the center of gravity of the distribution. Statistically, the mean is the 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 mode for nominal or ordinal data as simply the category that has the highest frequency. For interval or ratio data, the mode is the most frequently occurring value which is given by the Excel function MODE(). For continuous data with no repeated values, a modal range can be located on a histogram as the bin with the highest sample count. However, this mode will vary according to the choice of bin width and bin range used in the histogram. A widely-used numerical estimate of the mode can be made for a moderately asymmetrical distribution by the relationship:

Mode = 3*Median - 2*Mean

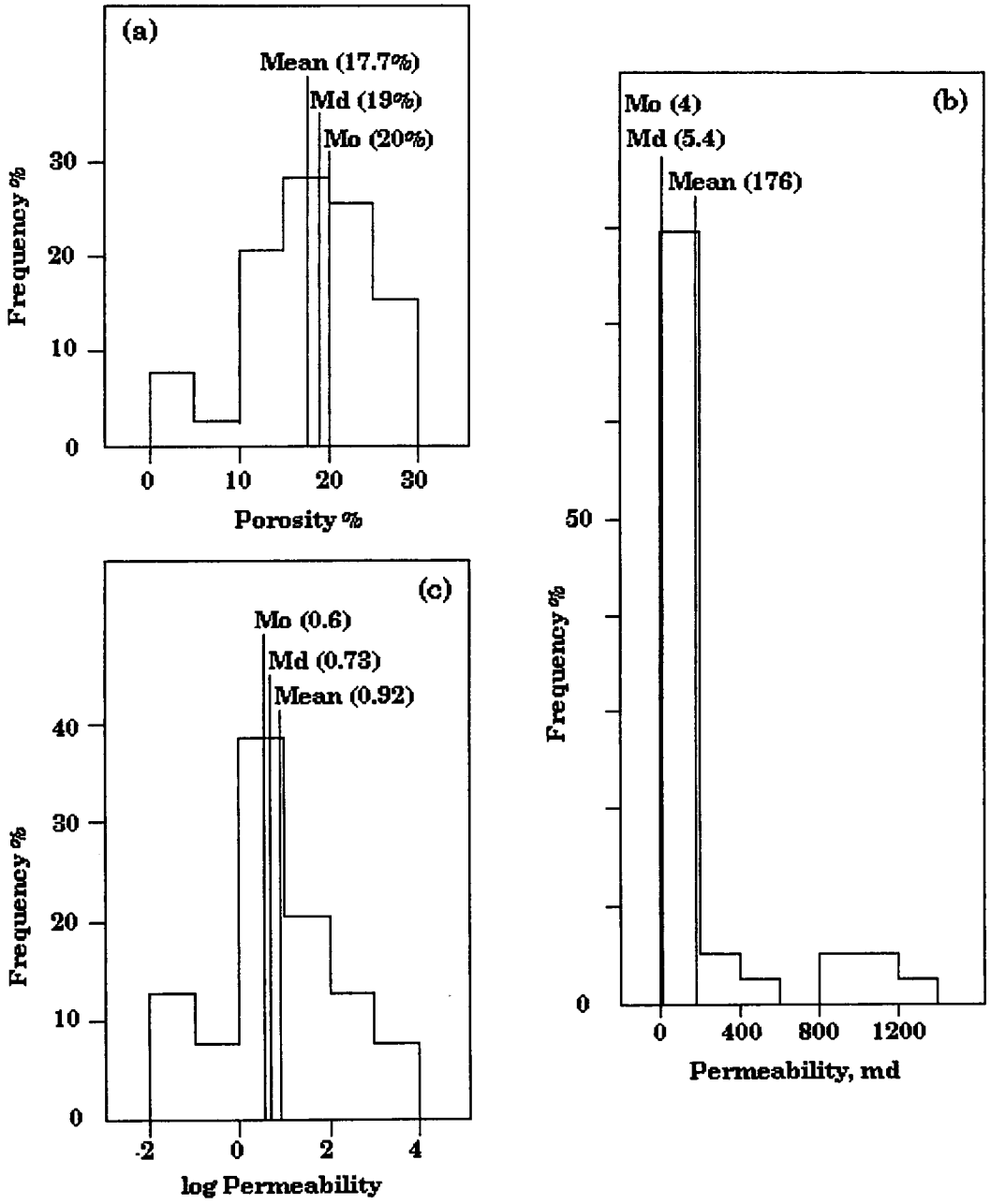


TORIS database: examples of central measures of sandstone oilfields



TORIS database: comparison of sandstone oilfields porosities by geologic age.

The three central measures are shown for porosity and permeability distributions in the Picaroon sandstones. 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: $\overline{X}g = \sqrt[n]{\Pi X_i}$.



Different averages: the arithmetic, geometric, and harmonic means

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

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

where n is the number of observations.

The **geometric mean** is the nth root of the product of the n observations and so is a multiplicative average calculated by:

$$\overline{X}g = \sqrt[n]{\Pi X_i} .$$

This is equivalent to the antilog of the arithmetic average of the logarithms of the observation values.

$$\overline{X}_g = 10^{\left(\sum \log_{10} X_i\right)/n}$$

The **harmonic mean** is the reciprocal of the arithmetic average of the reciprocals of the observation values:

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

The arithmetic mean is most commonly used. It is appropriate for physical models where the deviations of observations from a central value can be considered as the summation of small arithmetic displacements. The sample estimate of this mean is adversely affected by anomalous extreme values ('outliers').

The location of the geometric mean matches that of the arithmetic mean of observations that are transformed to a logarithmic scale. It is appropriate for physical models that are controlled by multiplicative processes. The sample estimate of the geometric mean is less sensitive to observations with high values.

The harmonic mean is used for averaging rates or ratios. It is insensitive to high observation values and controlled by observations with low values.

For any data set, the arithmetic mean is greater than the geometric mean, which is greater than the harmonic mean or:

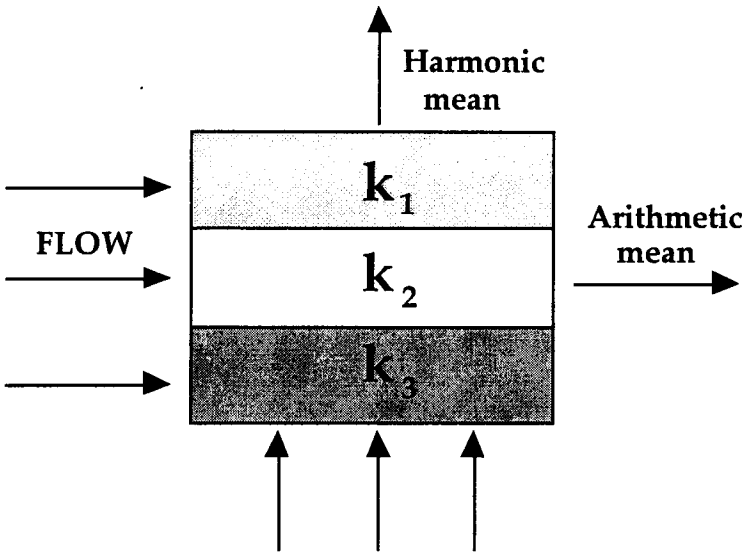
$$\overline{X}_h < \overline{X}_g < \overline{X}_a$$

So, for example, the mean values for permeability in the Picaroon Sandstone are:

arithmetic mean = 176 md
geometric mean = 8.2 md
harmonic mean = 0.18 md

Averaging permeabilities: arithmetic, geometric, harmonic means and the power law scaling method

The effective permeability of a layered medium is computed from an average of the permeabilities of the layers. However, the average depends on the direction of flow with respect to the layers, so when flow is parallel to the layers, the effective permeability is the arithmetic mean. When the flow direction is normal to the layers then the permeabilities are arranged in series rather than parallel and the effective permeability is the harmonic mean.



The two means provide two extremes of effective permeability, with the arithmetic mean as the upper bound and the harmonic mean as the lower bound. The "real" effective permeability generally lies between these two extremes. If there is no preferential direction of flow with respect to permeability variation, then the effective permeability is estimated by the geometric mean of the permeabilities. The three alternative means are considered to be special cases of a generalized power law of scaling (Desbarats, 1989), given by the equation:

$$\bar{k} = \left[\frac{1}{n} \sum_{i=1}^{n} k_i^p \right]^{\frac{1}{p}}$$

where p is an averaging exponent ranging between -1 (harmonic) and 1 (arithmetic). As p approaches zero, the effective permeability converges on the geometric mean. The power-law equation allows any intermediate averaging to be made, although the criterion to select a value for p is unclear (Jensen et al, 2000, p. 121).

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

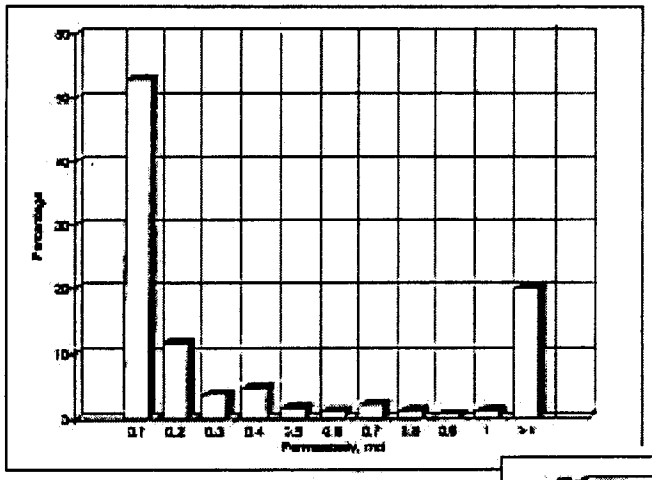
The use of the arithmetic average would disqualify all four formations; the median would designate all four as tight gas formations. In reality, large acreage blocks were selectively excluded, so that the permeability of the remaining acreage would have an arithmetic average of less than 0.1 md.

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:

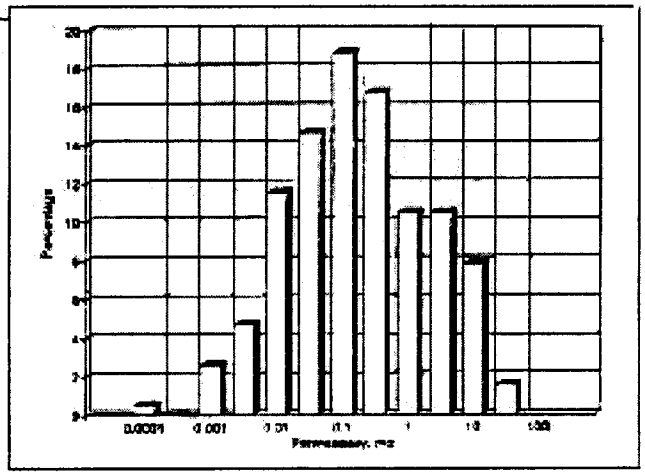
	CASE 1	CASE 2
Arithmetic mean	1.350 md -- 5.64 Bcf	0.284 md -- 5.30 Bcf/well
Geometric mean	0.106 md -- 4.00 Bcf	0.017 md -- 1.26 Bcf/well
Median	0.085 md -- 3.62 Bcf	0.020 md -- 1.41 Bcf/well
ACTUAL	**3.42 Bcf**	**1.75 Bcf/well**

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. 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.



Permeability distribution in Travis Peak



Log permeability distribution in Travis Peak

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.

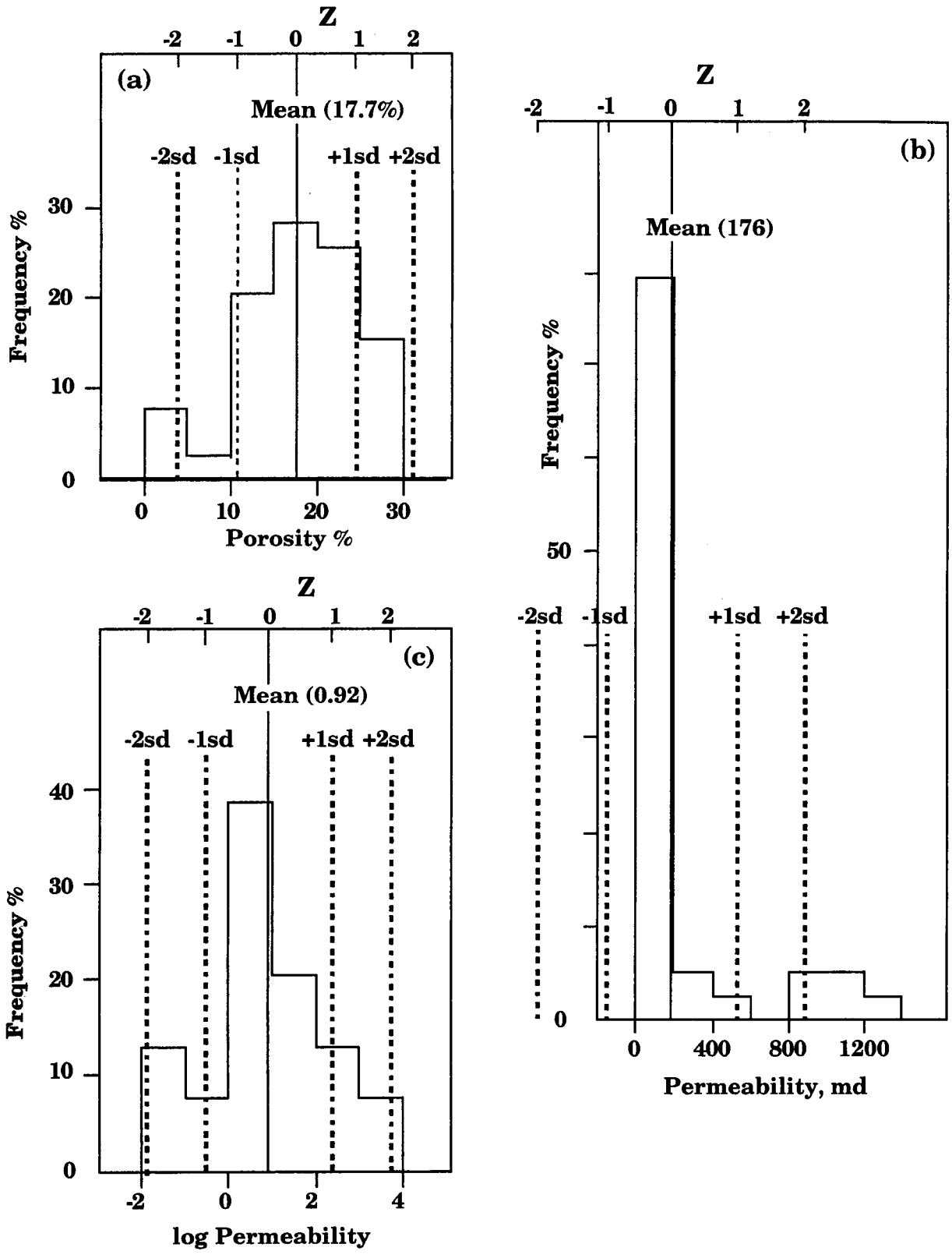
Variance is expressed in squared units, so a more tractable measure of spread is given by the standard deviation, s, which is the square root of the variance. The spread of a distribution can then be summarized as multiples of standard deviation distances about the mean value. This leads to a useful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

A dimensionless measure of variation that is commonly used is the coefficient of variation which is the standard deviation divided by the mean: $C_v = \frac{s_X}{\overline{X}}$

Plots of the standard deviation ranges of porosity, permeability, and logarithmic permeability of the Picaroon sandstones illustrate these concepts. 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:

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

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

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

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

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

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.

positive (right) skew

negative (left) skew

The fourth moment is kurtosis, given by:

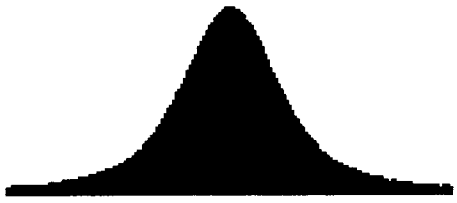
$$m_4 = \frac{\Sigma(X_i - \bar{X})^4}{(n-1)}$$

with its dimensionless equivalent calculated by:

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

The kurtosis measures the relative "peakedness" of the distribution. More peaked distributions are "leptokurtic"; flatter distributions are "platykurtic".

platykurtic



leptokurtic

More realistically, the useful information is generally reflected in the corresponding "fattening" or "thinning" 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= -0.60	Kt= 0.13
Permeability:	Sk= 2.10	Kt= 3.05

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 *expectations* for dimensionless values of skewness and kurtosis for a normal distribution are zero. Zero will occur for the normal *population*, but small non-zero numbers are inevitable for samples from a normal distribution. So, if we use the normal distribution as a reference curve, what are the minimum values of skewness and kurtosis that are to be exceeded before we can make a statement that the observed distribution is skewed (positive or negative) or that it is leptokurtic or platykurtic? The critical values will be controlled by the sample size and the probability value that we assign to our decision as to whether the statistic reflects normality or non-normality.
The standard deviation of the sample estimates of skewness for sample sizes of n observations taken from a normal distribution is the standard error of the skewness, s_{es} and is calculated as:

$$s_{es} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

At large sample sizes, this standard error is often approximated as SQRT(6/n). For a normal distribution, 95% of the values are within about (1.96) two standard deviations from the mean. So, if the calculated value of skewness is more than twice the standard error distance, we are outside the 95% confidence limits and (at that probability level) characterize the distribution as positive or negative skewed according to the sign of the statistic.
The standard deviation of the sample estimates of kurtosis for sample sizes of n observations taken from a normal distribution is the standard error of the kurtosis, s_{ek} and is calculated as:

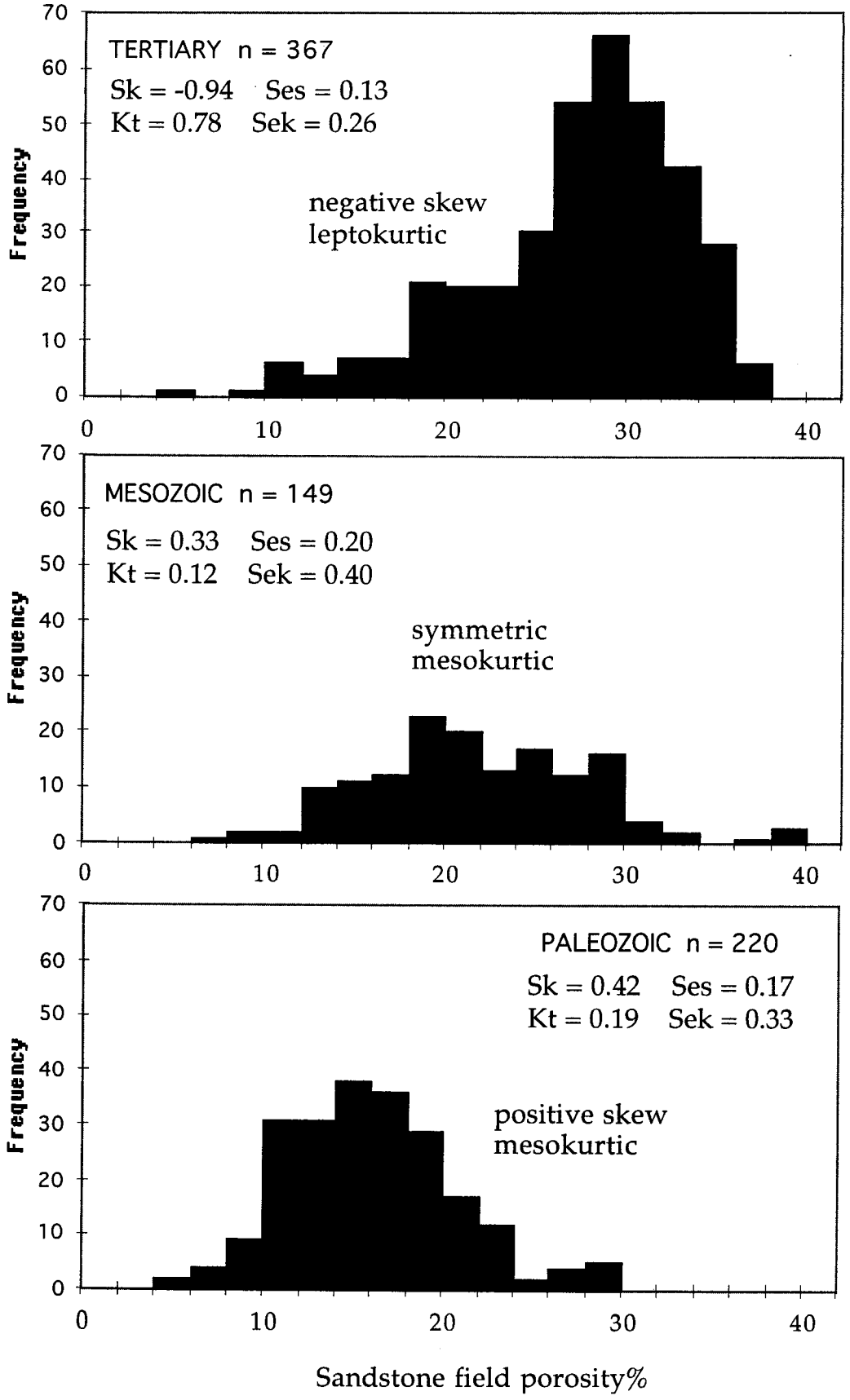
$$s_{ek} = \sqrt{\frac{4(n^2-1)s_{es}^2}{(n-3)(n+5)}}$$

At large sample sizes, this standard error is often approximated as SQRT(24/n). The development of a critical value for application to the sample statistic of kurtosis follows the same logic as for skewness.

The sample statistics of skewness and kurtosis for TORIS database porosities of oilfields in Tertiary, Mesozoic, and Paleozoic sandstone reservoirs are shown and compared with their corresponding critical test values, which are then used to characterize the distribution shapes either as:

negatively skewed, symmetrical, or positively skewed

and platykurtic, mesokurtic, or leptokurtic.



TORIS database: comparison of sandstone oilfields porosities by geologic age.

A comparison of the skewness and kurtosis of the TORIS database sandstone oilfields by geologic age shows characteristics that are consistent with the petrophysics of sand reservoirs. Tertiary sandstones are often poorly cemented and porosity values approach the maximum that is physically possible for a pack of sand grains. This physical constraint truncates a potential upper porosity tail to balance lower porosity sandstone reservoirs where pore space has been occluded by cements and clays (such as the Picaroon Sandstone). The end result is a leptokurtic distribution with negative skew. The Mesozoic sandstone oilfields are symmetric and mesokurtic because the skewness and kurtosis statistics are considered not significantly different from a normal distribution expectation when using a 95% probability criterion. However, notice that both statistics are positive, so the Mesozoic oilfield porosities are slightly positively skewed and mildly leptokurtic (think: "fat tails"). The Paleozoic oilfield sandstone porosities have a positive skew and are mesokurtic. At the lower porosity range, there is the mathematical constraint that porosities cannot be less than zero. This will contribute to the asymmetry of the distribution with a skew to higher values.

In this TORIS database example, the moments of skewness and kurtosis have been applied as useful descriptors of shape, with implications of physical causative processes linked with sandstone reservoir development. However, these same measures are commonly used in inferential statistics as a check on whether an observed sample distribution can be considered to be approximately normal. If the distribution is decidedly non-normal, then the standard methods of parametric statistics are compromised badly in their application to such data, with a potential for misleading or nonsensical results. This is because parametric methods rely entirely on the parameters of means and variances (and covariances) in analysis and these parameters mathematically define the univariate or multivariate normal distribution.

THE NORMAL DISTRIBUTION

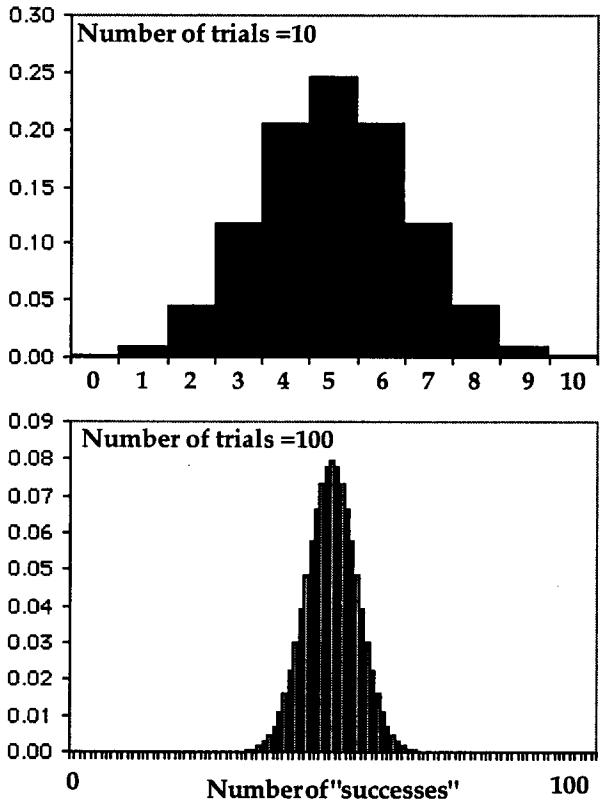
The normal distribution is the bell curve that is familiar even to those who know little or nothing about statistics. What is it? Where did it come from? Why is it important? The normal distribution is the limiting continuous form of the discrete binomial distribution. The binomial distribution applies 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. When the probability p = q = 0.5, then the binomial distribution for ten trials and 100 trials are:

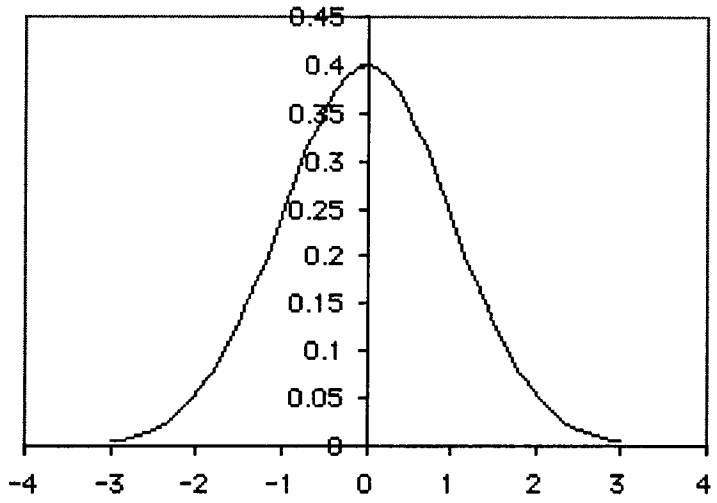


and are generated easily, using the EXCEL function BINOMDIST(s,n,p,c), where s is the number of successes, n is the number of trials, p is the probability of success, and c is FALSE. The use of FALSE generates the pdf (probability density function); TRUE generates the cdf (cumulative density function).

As n becomes larger, the distribution grows progressively smoother until in the continuous limit (an infinite number of trials), it becomes the normal distribution. 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 equation is used in the EXCEL function NORMDIST(x, mean, standard deviation,FALSE). The use of FALSE generates the pdf (probability density function); TRUE generates the cdf (cumulative density function).



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:

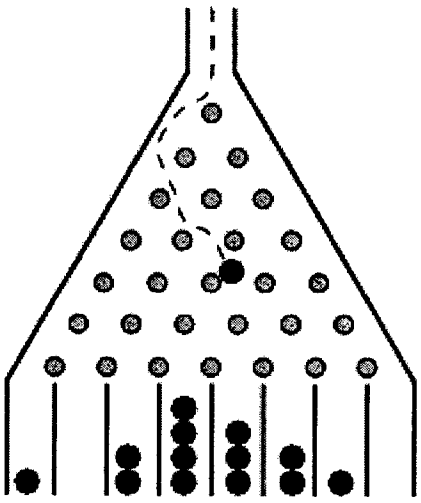
$$Z = \frac{(X_i - \overline{X})}{s} \text{ for a sample, or } Z = \frac{(X_i - \mu)}{\sigma} \text{ for a population.}$$

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. Most statistics text contain a normal distribution function table which relates Z-scores to the corresponding proportional area under the curve. Later, we shall see how this information is used in statistical inference tests.

It was the great German mathematician, Carl Friedrich 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 can also be demonstrated with a device known as a "Galton board" or a quincunx, invented by Francis Galton:

Galton Board (Quincunx)



Pascal's triangle

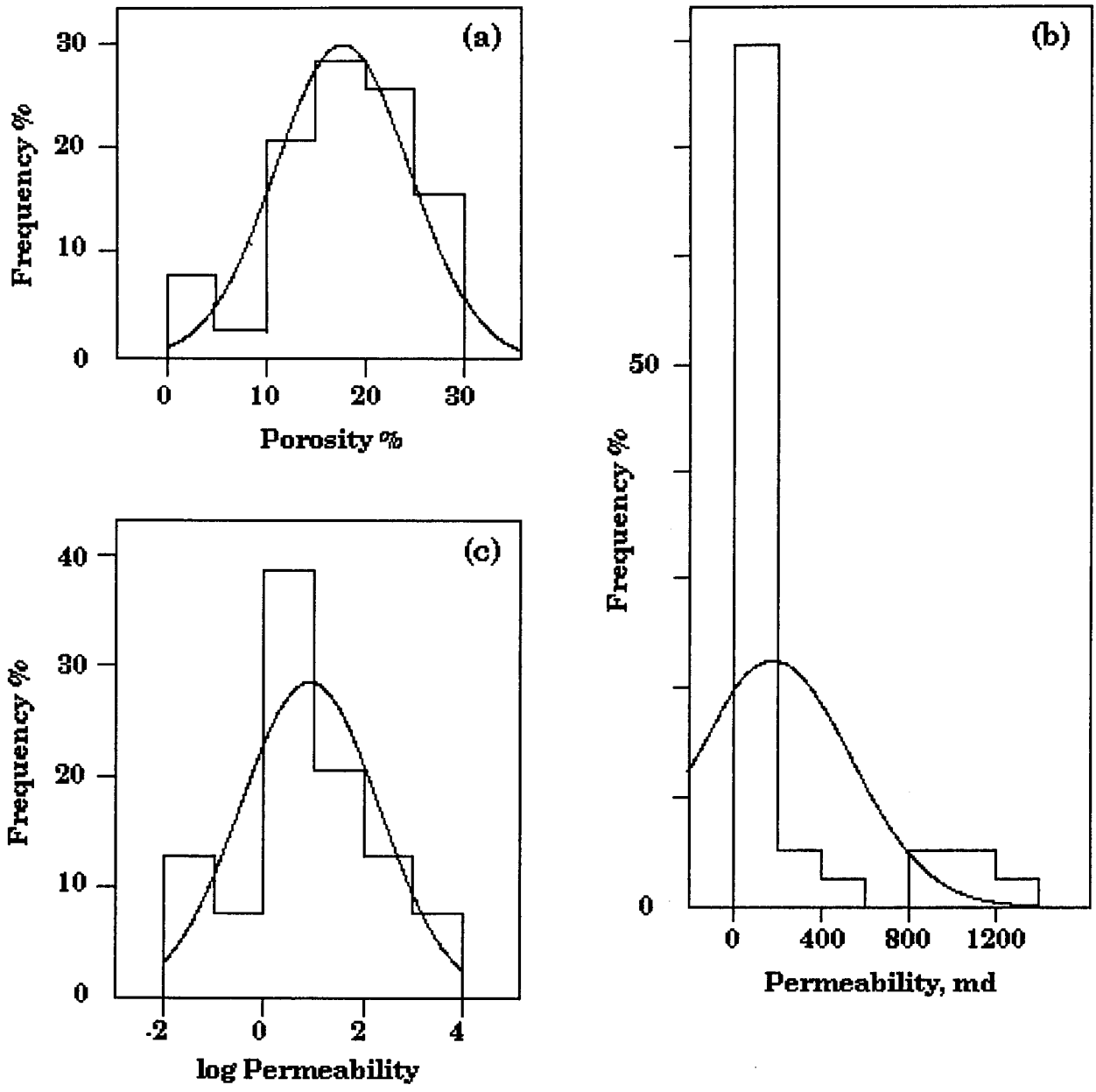
```
          1
        1   1
      1   2   1
    1   3   3   1
  1   4   6   4   1
1   5  10  10   5   1
1   6  15  20  15   6   1
1   7  21  35  35  21   7   1
```

With an equal chance of deflecting in either direction when a falling ball hits a pin, the distribution of balls that accumulate below follows a binomial distribution. A normal distribution would be approximated by a giant Galton board. Notice that the numbering of alternative pathways is also given by Pascal's triangle, which is another realization of the binomial distribution.

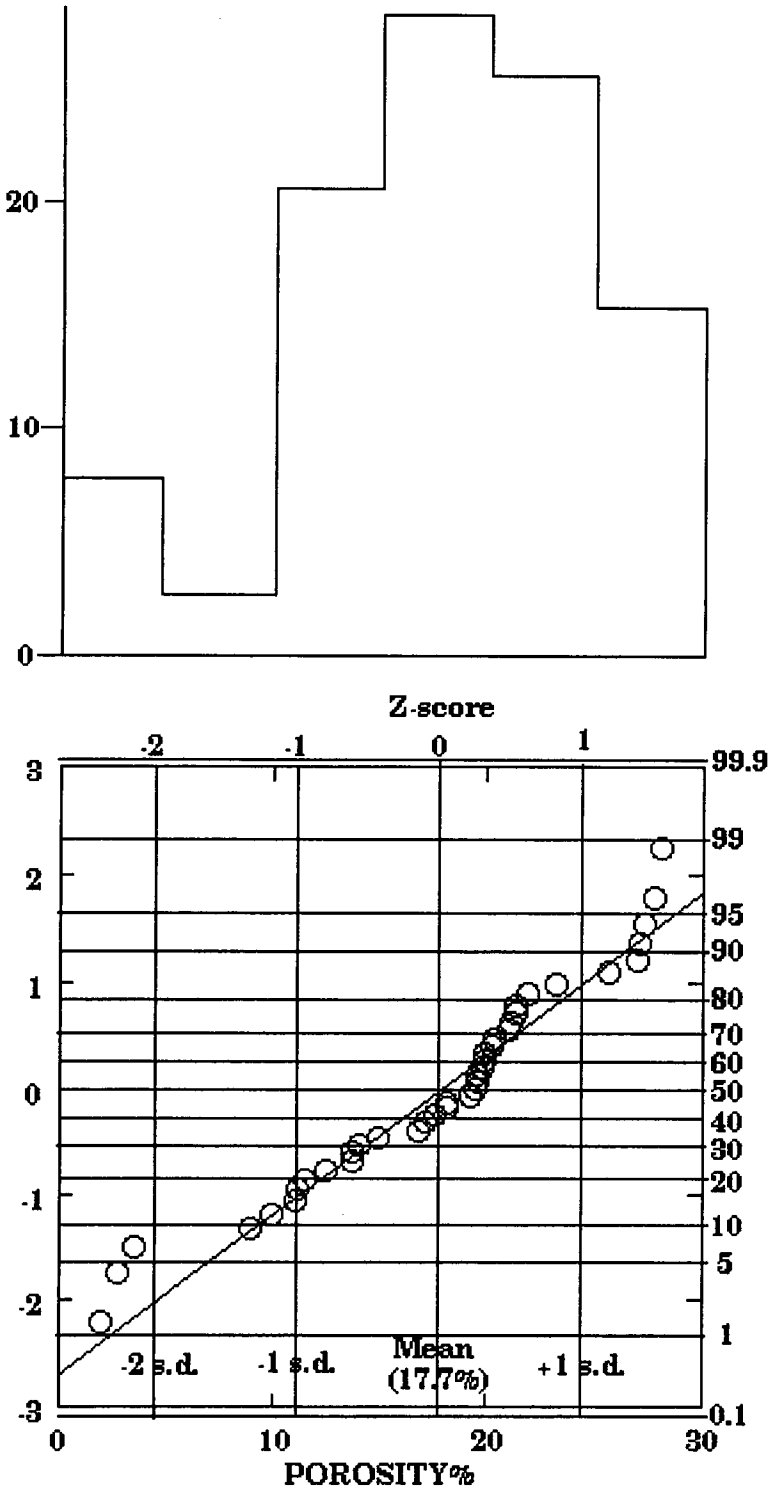
The normal distribution was proposed as a theoretical model for the analysis of statistical error (and still is used for that purpose). However, natural variabilities (height, weight, porosity) often took the form of a more-or-less symmetrical clump of values that faded into tails of less frequent and 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. 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 value. If normal, the points should plot on a straight line. P-plots are also available for lognormal and other hypothetical distributions.



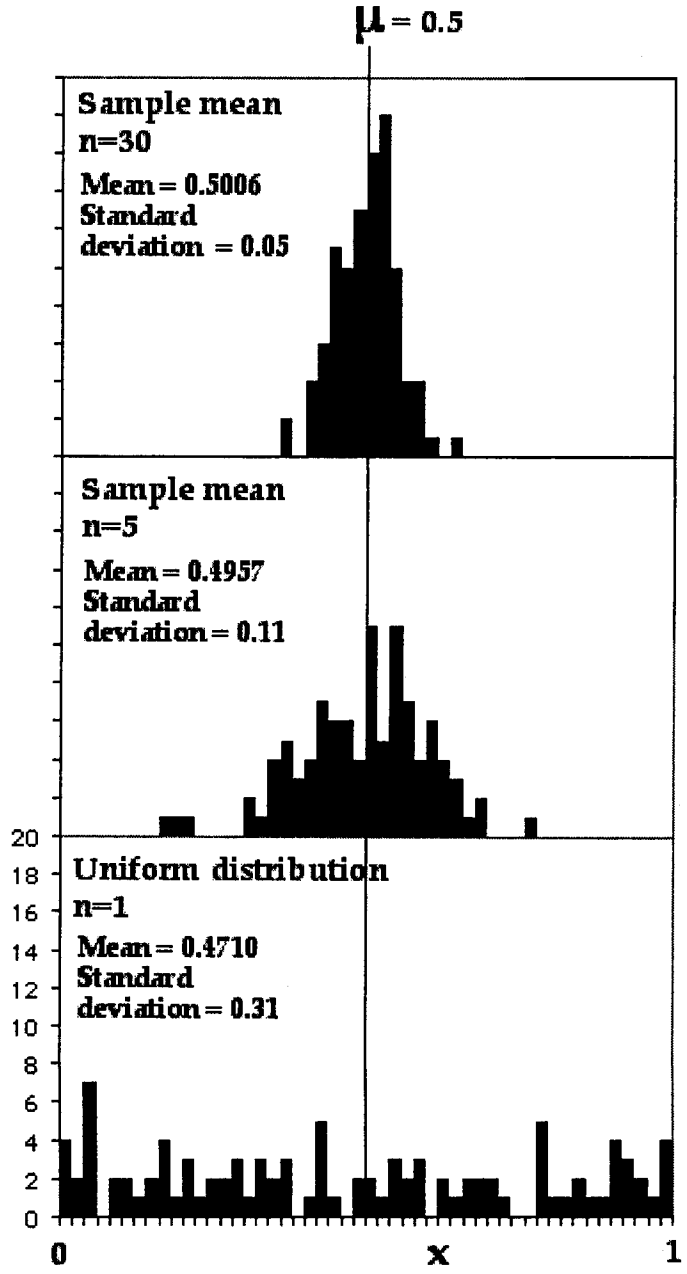
Normal distribution curves fitted to Picaroon Sandstone histograms of (a) porosity, (b) permeability, and (c) logarithmically-scaled permeability.



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. In the example below, the distribution of values uniformly distributed between 0 and 1 (generated by EXCEL's RAND()) is compared with the distribution of means of these values for sample sizes of 5 and 30.



The Central Limit 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:

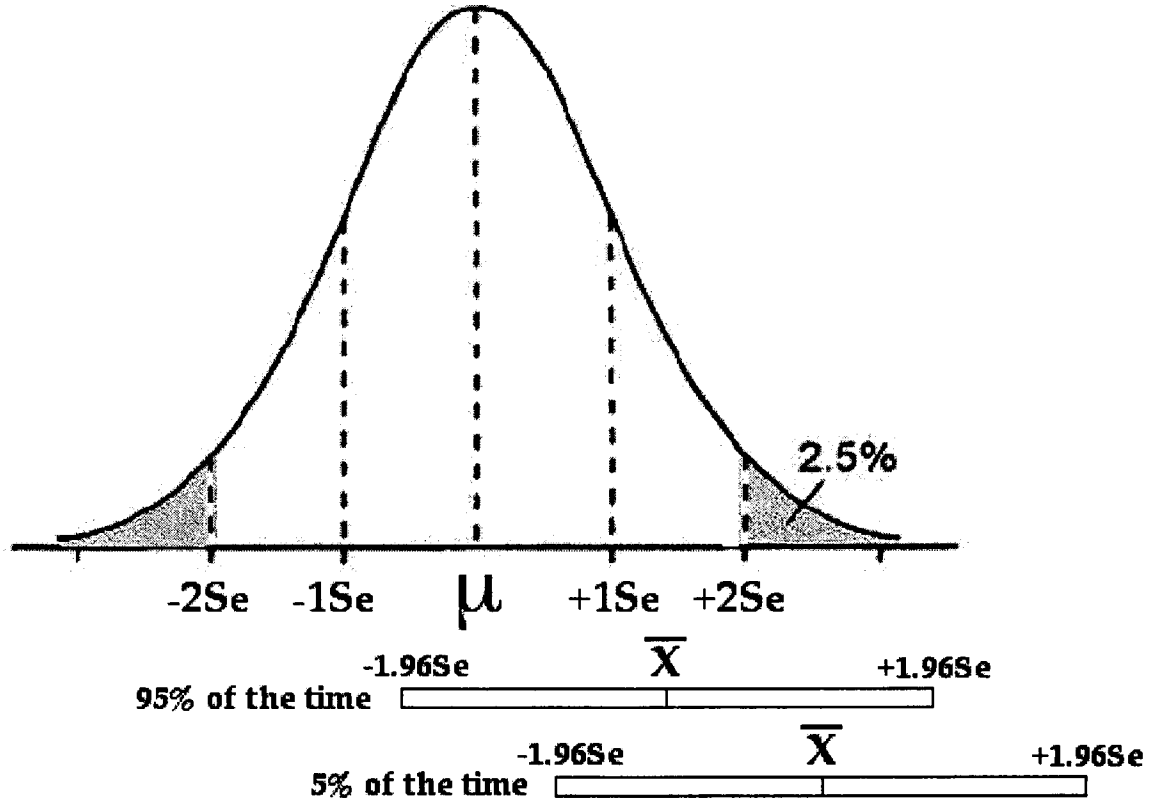
$$s_e = \frac{6.7}{\sqrt{39}} = 1.07$$

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%.



Let us next suppose that the management of the company has declared this result to be unacceptable: they need to have a 95% 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 the Picaroon Sandstone we would need to satisfy this demand. The required standard error is:

$$s_e = \frac{1}{1.96} = 0.51\%$$

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

$$n = \left(\frac{s}{s_e}\right)^2 = \left(\frac{6.7}{0.51}\right)^2 = 172.6$$

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:

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

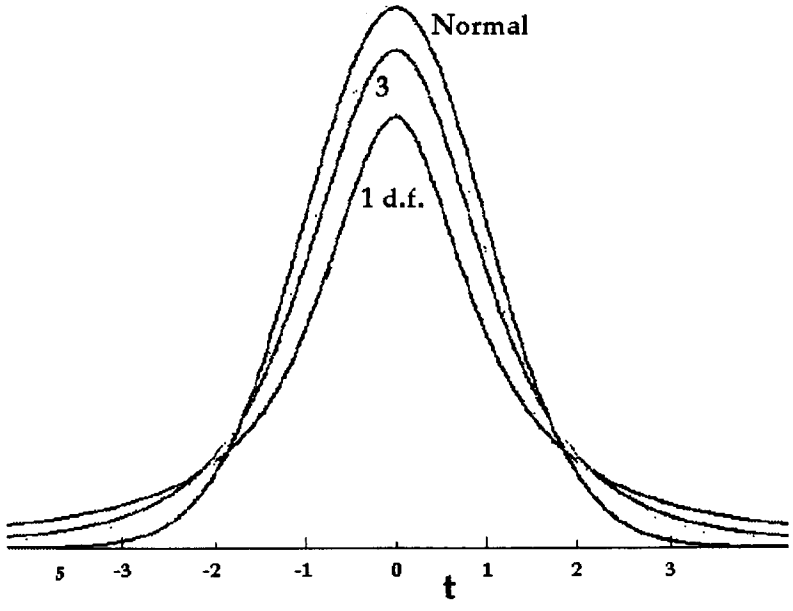
The standard error is given by:

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

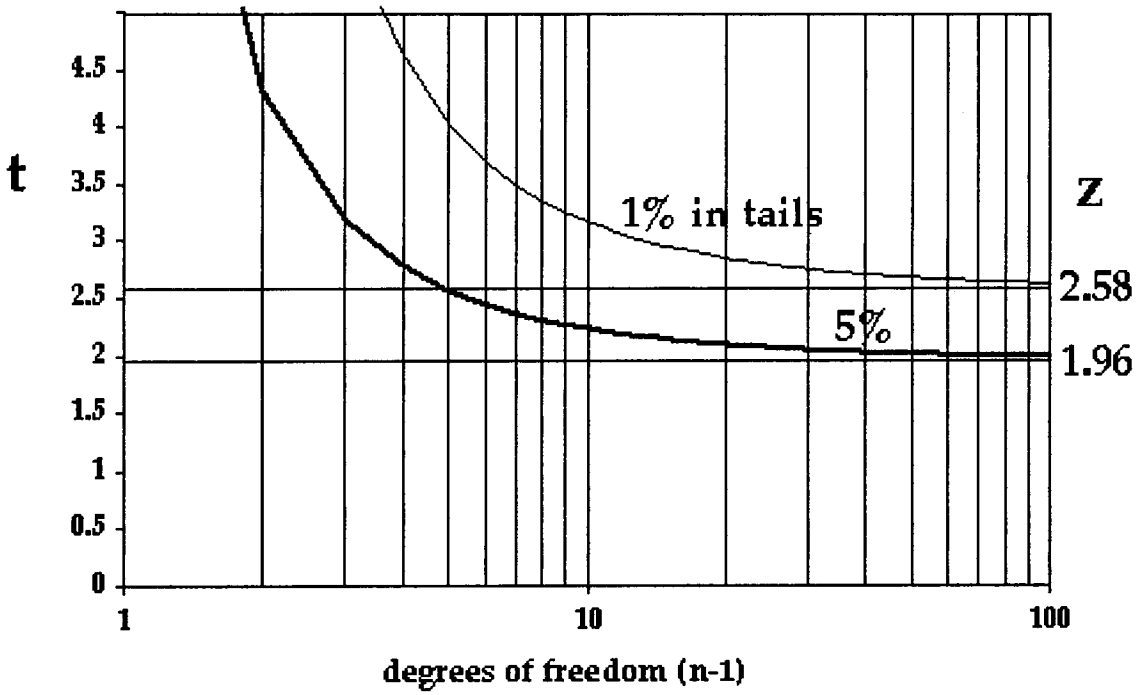
As we noted before, we generally do not know the population standard deviation, σ, so we substitute the sample estimate of *s* to give:

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

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*. Notice that the value of t is plotted in terms of the number of degrees of freedom, which is simply the sample size minus one (n-1). As *n* grows larger, the distribution contracts until it converges on the normal 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 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 graph above was created in EXCEL by using the t-distribution function of TINV(p,degrees of freedom) where p is the proportion of the t-distribution contained in the tails (e.g. 0.05 for 5%) and the output of the function is the corresponding value of t.

The t-distribution is the basis for the most common test of statistical inference: the **t-test** in which the difference between two sample means is evaluated with respect to the dispersion of observations about the means to determine whether the samples are likely to be taken from the same population or alternatively, different populations.

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

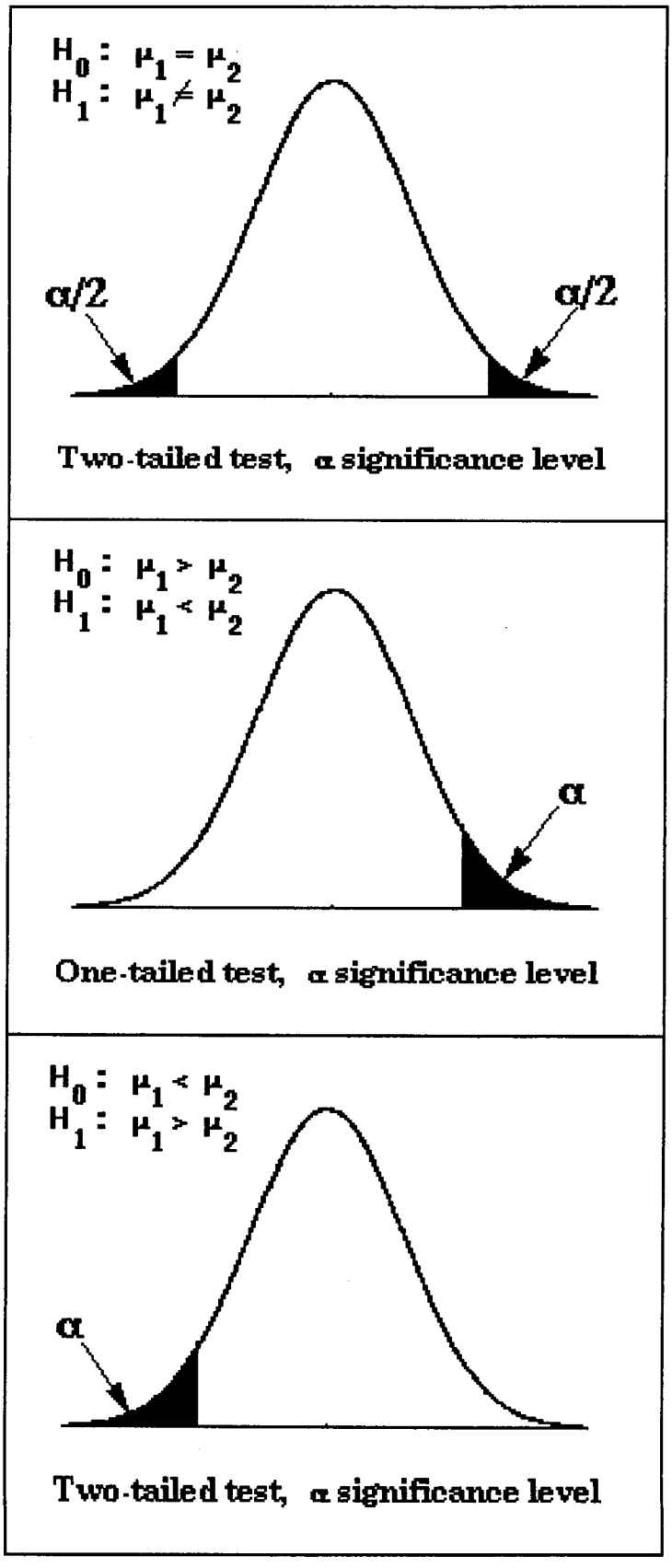
A significance level is the acceptable risk of making a Type I error: the probability that the null hypothesis is really true even when it was rejected by the statistical test. The significance level is denoted by α and should be selected *before* the test. This convention cuts down on gerrymandering 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. 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 BEGINNINGS OF STATISTICAL INFERENCE

Ronald A. Fisher (1890-1962) was the pioneer of statistical inference who proposed that most research projects were exercises in inductive logic. The results of a few controlled experiments could be extended to generalizations concerning the phenomenon studied. Going from the observations recorded (the sample) to general statements (the population) involved a degree of uncertainty. However, the degree of uncertainty could be calculated by using statistical methods rooted in probability theory. The analysis could be controlled by the researcher through the design of the experiment before the observations were made and the application of statistical sampling procedures. Fisher devised the method of Analysis of Variance (ANOVA) and published his ideas in the classic book "The Design of Experiments" (1935) which started a revolution in research methods and the analysis of experimental results. Fisher's book starts with the famous illustrative example of the "tea-tasting experiment". It is supposed that there is a lady who claims she can tell the difference between a cup of tea in which milk has been added *after* the tea from a cup in which milk has been poured *before* the tea. Eight cups of tea are prepared, four one way, four the other, and presented to her in random order. She is further told that there are four cups of each type. If thew lady identifies all eight cups correctly, it is by no means certain that she can tell the difference between the two types of tea. . There is a 1/70 probability that this outcome could arise by chance alone. Therefore, if the null hypothesis that she cannot discriminate is rejected, there is a 0.014 chance that this conclusion is incorrect.

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**. In this case, the degrees of freedom would be given by:

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

and the standard error would be 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:

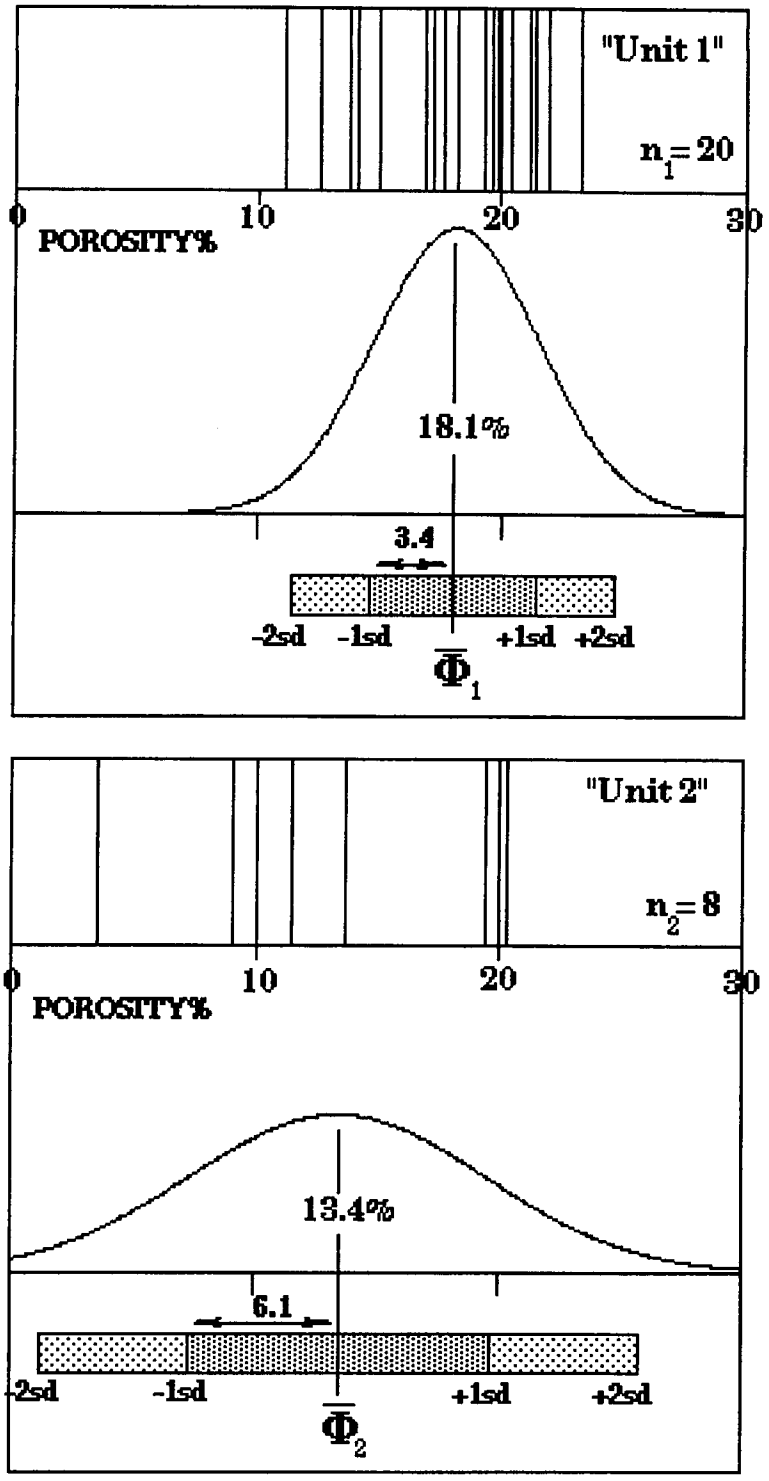
$$s_e = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

However, 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:

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



Comparison of prosity distribution between Picaroon Sandstone "Unit 1" and "Unit 2"

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 a 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 5%, then the α-level is 0.05. This decision means that we have accepted the contingency that if we reject the null hypothesis, then there is a one in twenty chance that we could be wrong. The critical value of t is found from the t-table as:

Critical t value @ α 0.05 and 9 df = 2.26

(This critical value is for a *two-tailed test* in which the five percent is allocated between the two tails of 2.5% each.)

The computed statistic does not exceed this critical value and so we do **not** reject the null hypothesis and accept the alternative hypothesis that they are samples taken from the same population. But visually, the porosities of Units 1 and 2 appear to be significantly different. Does this t-test outcome "prove"the surprising result that they are samples from a common population? No, it does not. Instead, it shows that the evidence was not sufficiently overwhelming to rule out the null hypothesis. Notice that if we had elected to go with a 10% significance level, we would have rejected the null hypothesis because the critical value of t is lowered to 1.833. So, the important factors to consider are the relatively small samples available in conjunction with the conservative, but traditional, α -level of 0.05, as well as the strong difference in sample variances which caused the estimated degrees of freedom to plummet.

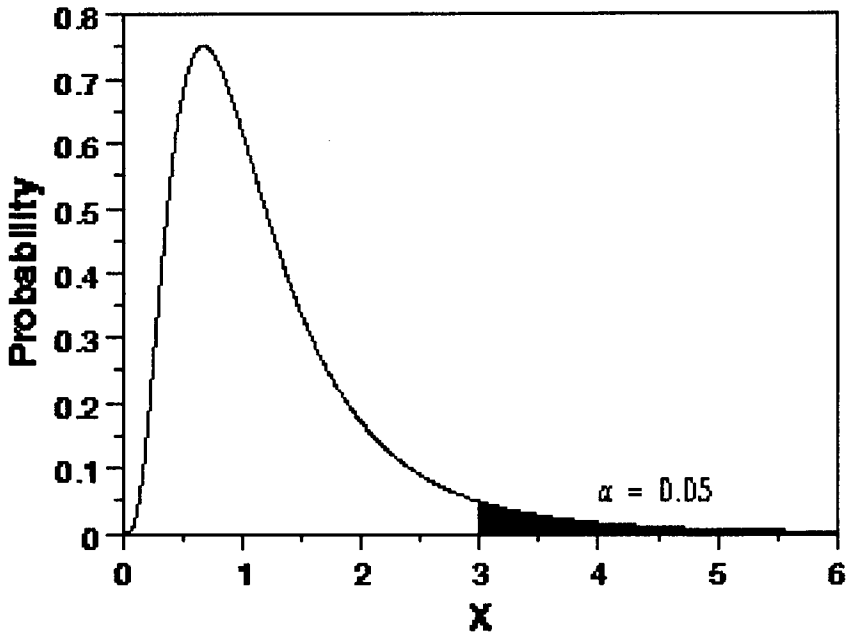
The same t-test results are shown as generated by EXCEL's Data Analysis Tools, when selecting the ALPHA value as 0.05 (the default value)::

t-Test: Two-Sample Assuming Unequal Variances		
	Unit1PHI%	*Unit2PHI%*
Mean	18.145	13.4375
Variance	11.38471	37.28554
Observations	20	8
Hypothesized Mean Difference	0	
df	9	
t Stat	2.05846	
P(T<=t) one-tail	0.034827	
t Critical one-tail	1.833114	
P(T<=t) two-tail	0.069654	
t Critical two-tail	2.262159	

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 shape of the F-distribution is asymmetric and the skewness varies with the degrees of freedom, becoming more symmetric at higher df numbers.

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

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

The alternative hypothesis is then:

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

The asymmetry of the distribution means that there will be different critical test values of F at each end of the distribution. So, the larger sample variance is placed as the numerator, the smaller as the denominator in the ratio computation. This changes the form of the test to a one-tailed test because the alternative hypothesis region is restricted to the upper tail because:

$$H_1: \sigma_1^2 > \sigma_2^2$$

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:

$$v_1 = n_1 - 1 = 7 \text{ and } v_2 = n_2 - 1 = 19$$

Then the degrees of freedom are used to locate the test value in an F-distribution 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. The same F-test results are shown as generated by EXCEL's Data Analysis Tools, when selecting the ALPHA value as 0.05 (the default value)::

F-Test Two-Sample for Variances		
	Unit2PHI%	*Unit1PHI%*
Mean	13.4375	18.145
Variance	37.28554	11.38471
Observations	8	20
df	7	19
F	3.275053	
P(F<=f) one-tail	0.018611	
F Critical one-tail	2.543537	

ANALYSIS OF VARIANCE (ANOVA)

The basic operation of analysis of variance (ANOVA) is to subdivide the total variability in a data set into components that can be identified with different sources. This is conventionally done by designing a sampling pattern such that individual observations can be combined in different ways to form composite observations. The null hypothesis is that the different variances are all sampled from the same population. If the null hypothesis is rejected, then the alternative is accepted that there are systematic components of variation within the data.

The most simple type of analysis of variance is a one-way design which is used to assess a single source of variation. The one-way ANOVA is basically an extension of the t-test of two sample means to allow comparison of multiple sample means.

Let us consider a situation where there are m groups, each of which has n observations. The grand mean is the mean of all the observations

$$\bar{\bar{X}} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}}{m \bullet n}$$

The total variation of the data is given by the sum of the squared deviations of the individual observations around the grand mean:

$$SS_T = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(X_{ij} - \bar{\bar{X}}\right)^2$$

This equation can be rewritten as :

$$SS_T = \sum_{j=1}^{m}\sum_{i=1}^{n}\left[\left(X_{ij} - \bar{X}_j\right) + \left(\bar{X}_j - \bar{\bar{X}}\right)\right]^2$$

where the group means have been added and subtracted to each term. If the equation is expanded, some of the terms will cancel, and the result is:

$$SS_T = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2 + \sum_{j=1}^{m}\left(\bar{X}_j - \bar{\bar{X}}\right)^2$$

This equation can be summarized as: $SS_T = SS_W + SS_B$ where SS_T is the total sums of squares, SS_W is the sums of squares within the groups, and SS_B is the sums of squares between the groups.

MS_B is basically the variance of the group means found by dividing the sums of squares between groups by the number of degrees of freedom which is the number of groups minus one: $MS_B = \frac{SS_B}{(m-1)}$

MS_W is a variance estimate based on the total variation from which the means have been subtracted: $MS_W = \frac{SS_W}{(n-m)}$

The two variances can be compared by an F-test. The F-test statistic is the ratio of the mean squares between the groups to the mean squares within the groups:

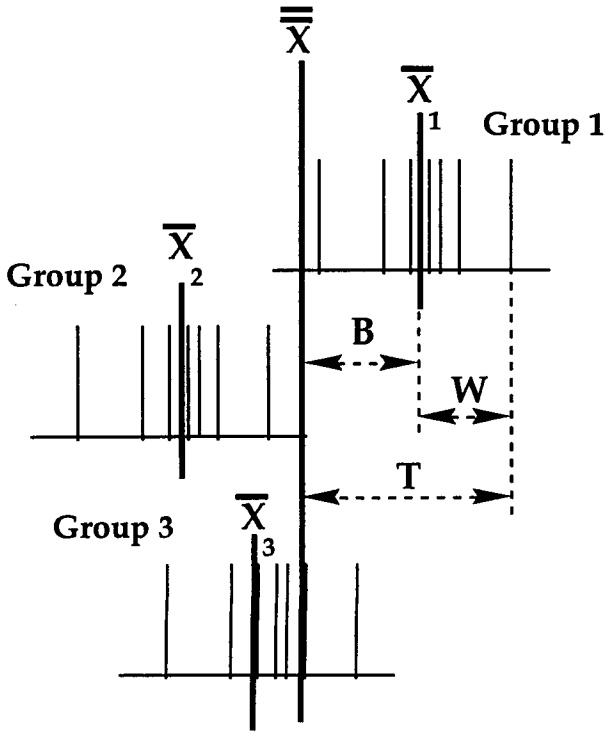
$$F = \frac{MS_B}{MS_W}$$

If they are found to be the same, then they both estimate the population variance and differences between the groups are not significant. If they are found to be different, then one or more groups is significantly different from the others.

The statistics of an analysis of variance are traditionally shown in an ANOVA table. The one-way analysis table takes the form:

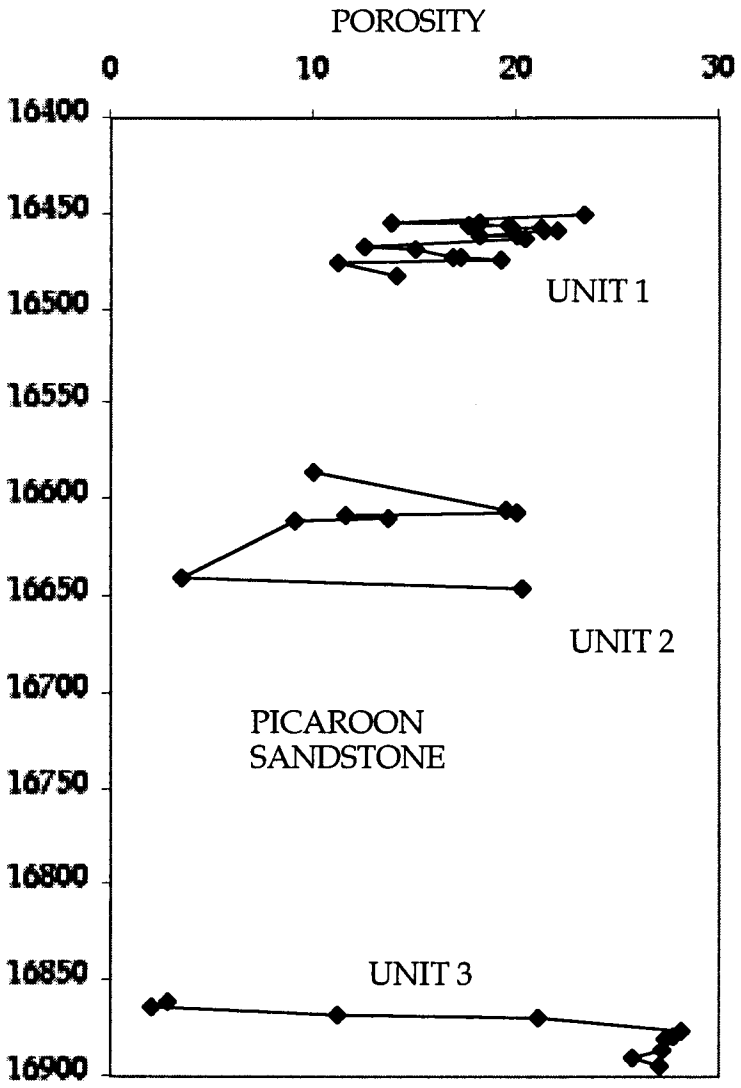
Source of variation	Sum of squares	Degrees of freedom	Mean squares	F-test
Between groups	$SS_B = \sum_{j=1}^{m}\left(\bar{X}_j - \bar{\bar{X}}\right)^2$	m-1	$MS_B = \frac{SS_B}{(m-1)}$	$F = \frac{MS_B}{MS_W}$
Within groups	$SS_W = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2$	n-m	$MS_W = \frac{SS_W}{(n-m)}$	
TOTAL	$SS_T = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(X_{ij} - \bar{\bar{X}}\right)^2$	n-1		

The F-statistic is compared with the critical F-value for the selected α-value (usually 0.05) and the two values of degrees of freedom of (m-1) and (n-m). If the F-statistic exceeds the critical value, then the null hypothesis of no difference is rejected.



Analysis of variance of potential differences in porosity between Picaroon Sandstone Units 1, 2, and 3

The variability of porosity within and between the "units" of the Picaroon Sandstone is shown graphically:



The EXCEL Data Analysis Tools has options for ANOVA. analysis. First, the porosities of Unit 1, Unit 2, and Unit 3 porosities are arranged in three columns (or three rows). The ANOVA: Single Factor option is selected and the results are shown below in the ANOVA table, where the ALPHA value chosen was 0.05 (the default value)::

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Unit1PHI%	20	362.9	18.145	11.3847105		
Unit2PHI%	8	107.5	13.4375	37.2855357		
Unit3PHI%	10	200.5	20.05	112.893889		
ANOVA						
urce of Variati	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Grou	204.460697	2	102.230349	2.39599184	0.10586879	3.26741656
Within Groups	1493.35325	35	42.6672357			
Total	1697.81395	37				

As with the t-test, the null hypothesis is not rejected at the α-value of 0.05, although EXCEL helpfully points out that rejection would have occurred at an α of 0.11. So, the statistics of porosity differences are not convincing from an orthodox statistical point of view, but might be sufficient from an engineering perspective, particularly when considering the small sample sizes and sampling pattern. What might the engineer have in mind when examining these results? Maybe the engineer would be considering whether or not there were sufficient grounds to subdivide the section into distinct flow units. But what if the units had been statistically different, but the difference in the porosities of the "statistical flow units" was trivial from an engineering perspective? .....

Two-way ANOVA and beyond

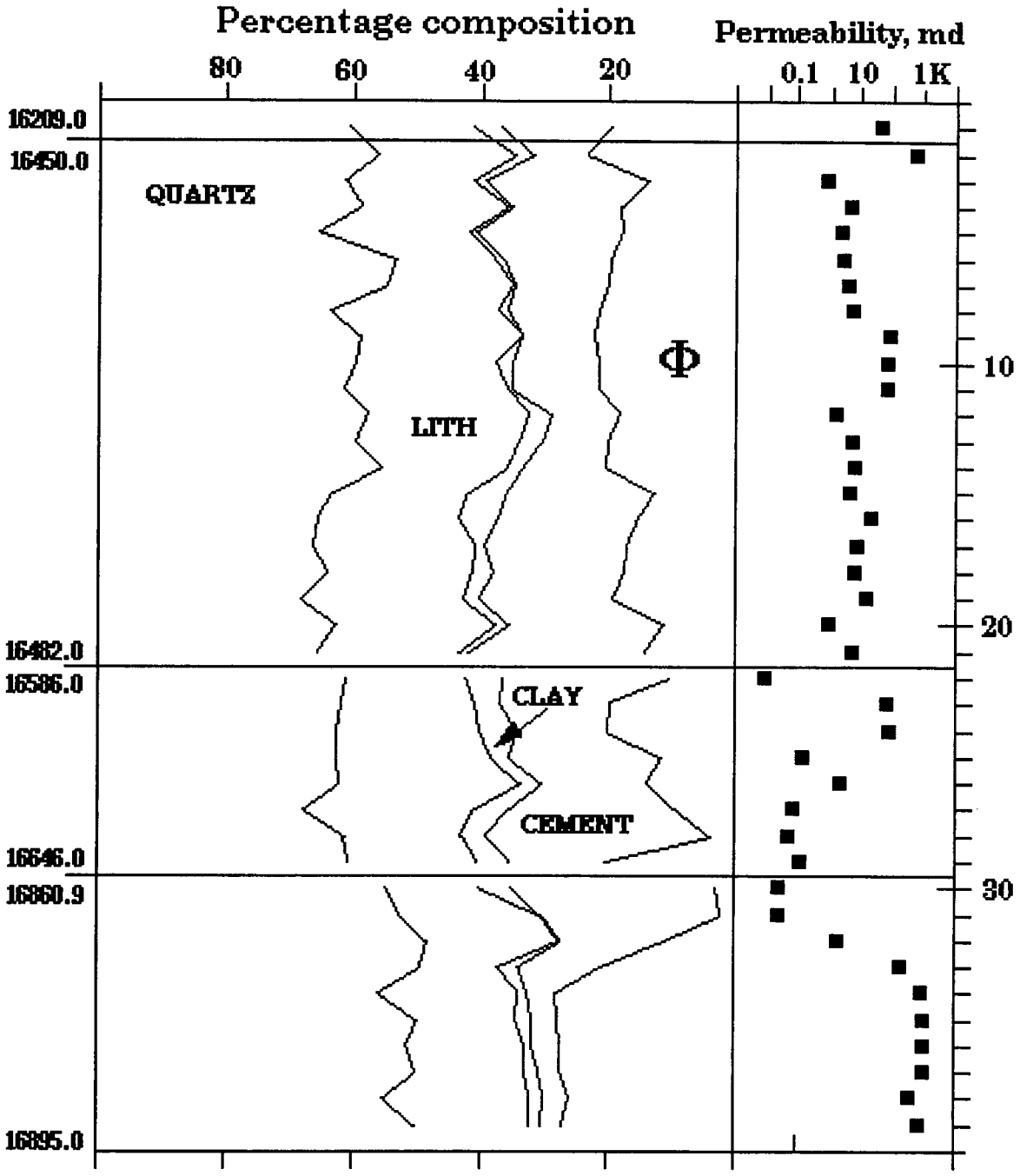
A two-way ANOVA can be constructed with rows representing differences with respect to one variable and columns with respect to a second variable. If more than one observation is located in a cell, then these multiple observations are termed 'replicates'. Differences can then be evaluated between rows, columns, and row-column effects that reveal interactions between the two variables. Even more complex designs can be created by nesting levels of variables or examining the effects of three variables simultaneously by a Latin-square ANOVA.

EXCEL has options for two-way ANOVA in the Data Analysis Tools. However, statisticians are critical of the performance of the EXCEL two-way ANOVA, with comments such as: " Excel should be used only for the most rudimentary of ANOVAs, the simple one-way between-subjects design (viz. independent t-test). For more complex designs, you should use a proper statistics program." The fact that ANOVA software packages are marketed as add-ins to EXCEL by companies outside Microsoft reinforces the validity of this negative opinion.

BIVARIATE ANALYSIS

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.



Composition-permeability profile of the Picaroon Sandstones

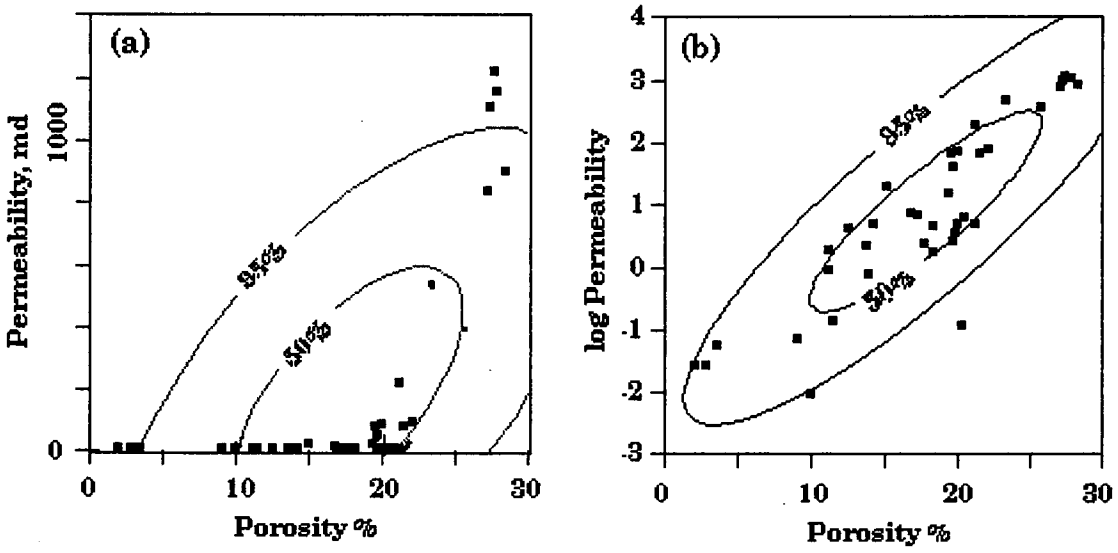
COVARIANCE AND CORRELATION

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:

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

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 but a marked improvement in (b) when the permeabilities are scaled logarithmically.



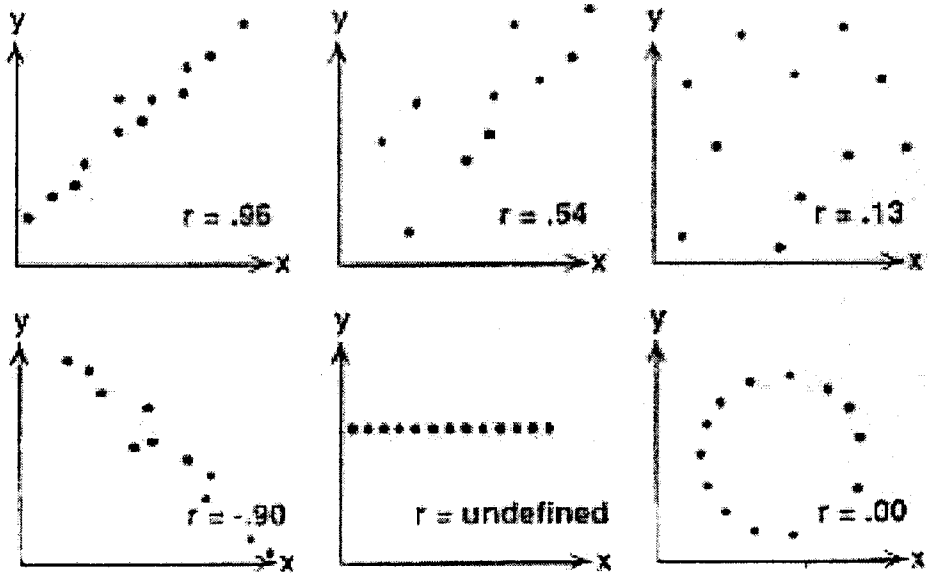
Bivariate normal distribution contours based on means, variances, and covariances fitted to (a) porosity-permeability plot and (b) porosity - logarithmic permeability plot of Picaroon Sandstone core measurements.

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:

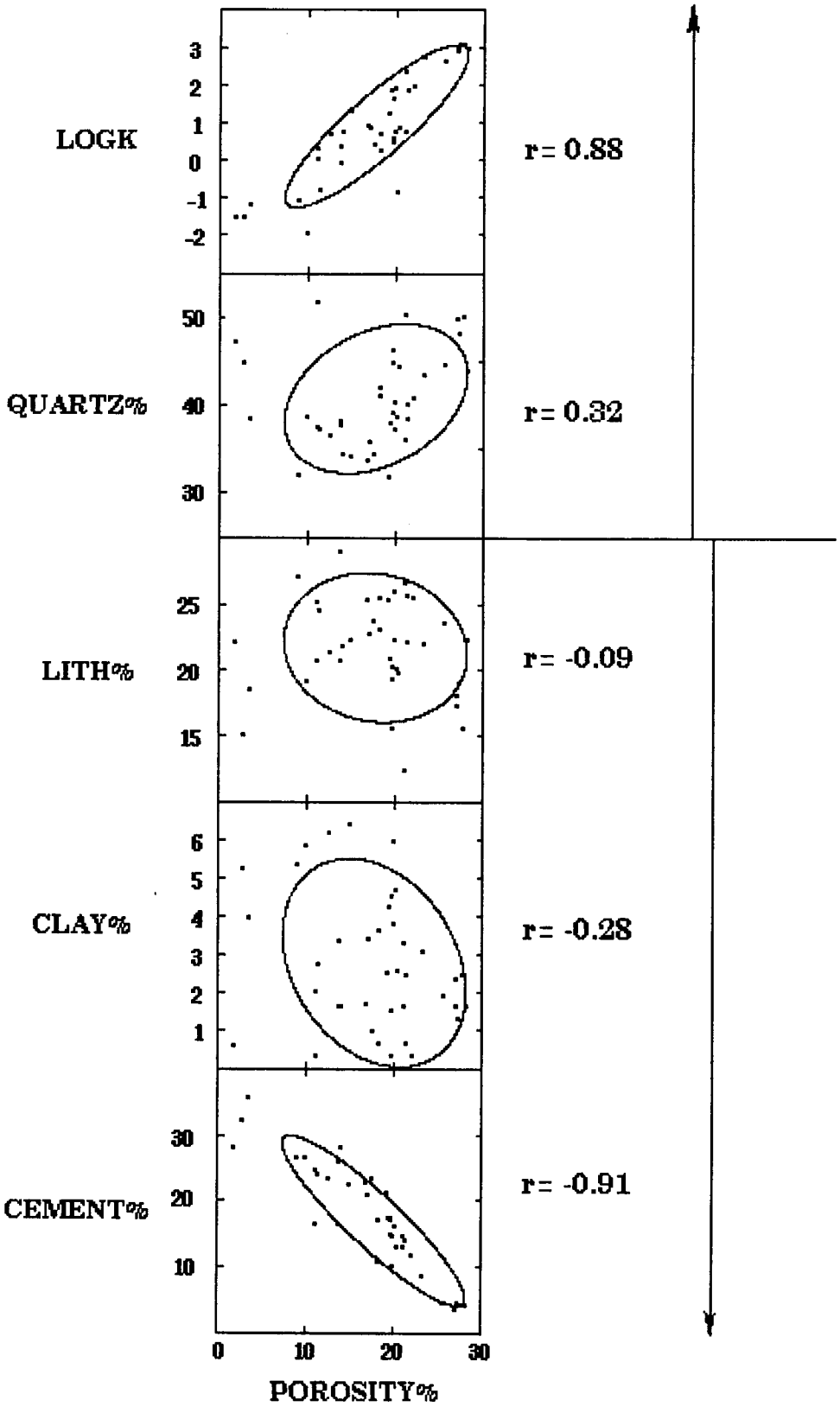
$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

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.



Crossplots, bivariate normal 90% contours , and correlation coefficients of Picaroon sandstone porosity correlations with other variables show interrelationships that match intuition. 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.



Picaroon Sandstone Pearson correlation coefficients of porosity versus logarithmic permeability and compositional content

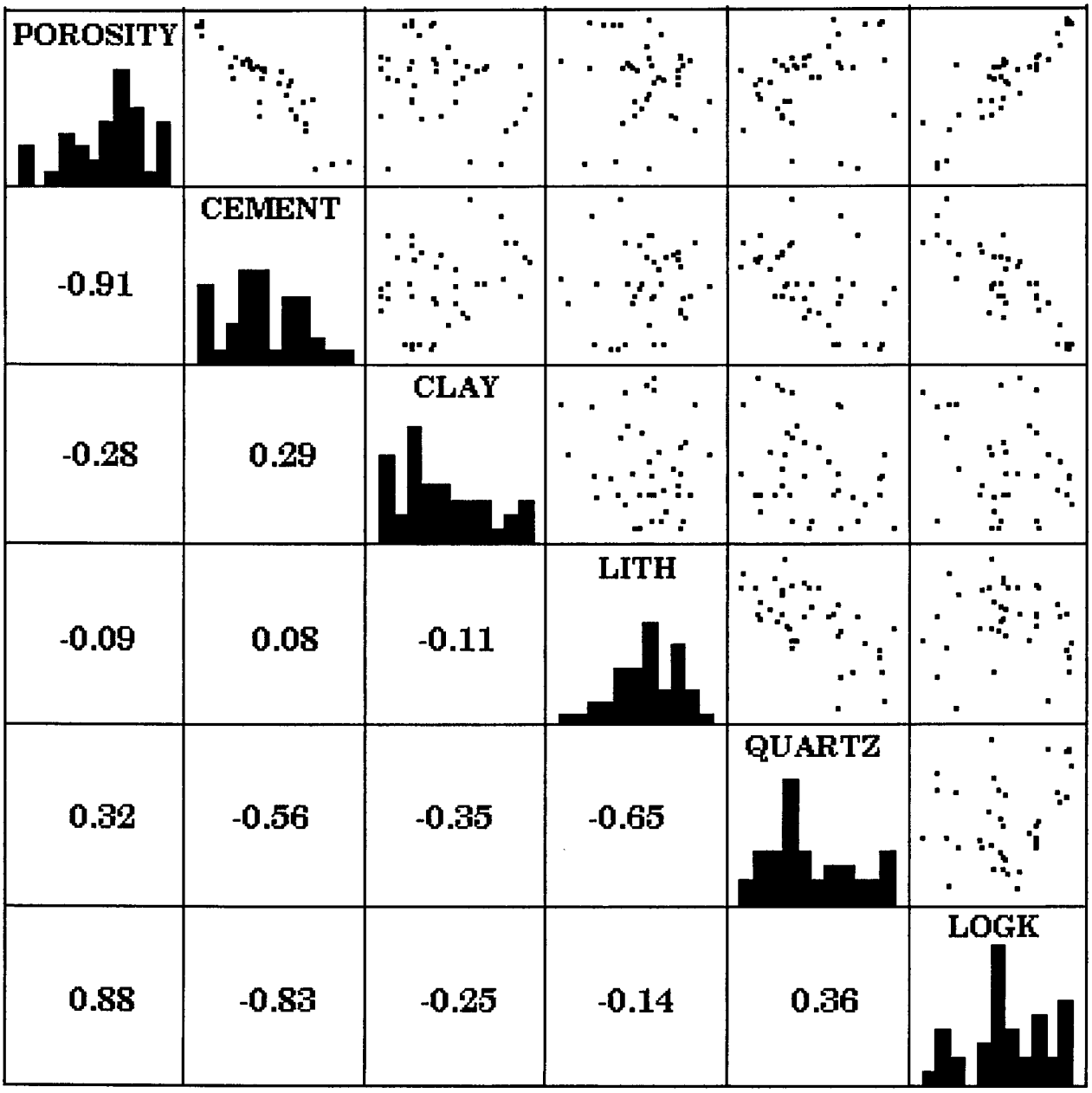
THE CORRELATION MATRIX

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).

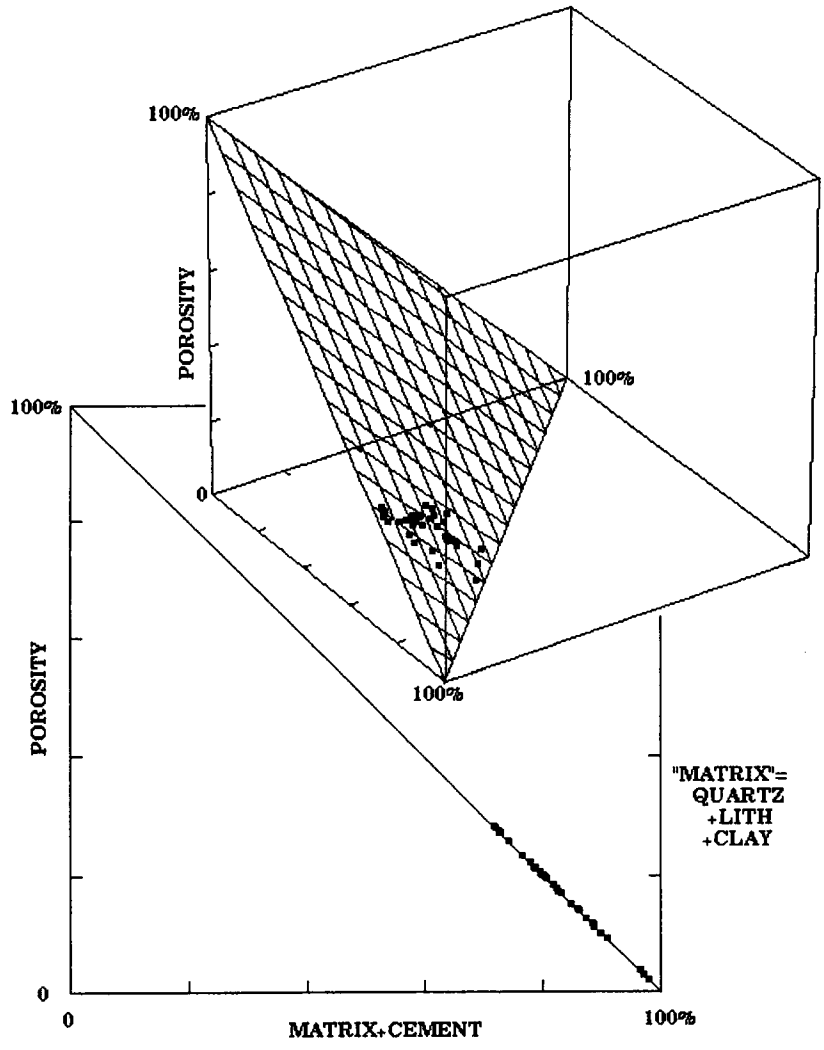
There are some other interesting correlations among the compositional variables. However, their interpretation is fraught with major difficulties, because the variables collectively make up a closed-system. As a result, much of the correlation structure is erroneous. The values and even the signs (positive 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 correlation coefficients of composition and logarithmically-scaled permeability

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. 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.



Closure in compositional variables illustrated by bivariate and trivariate condensations of Picaroon Sandstone data.

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;

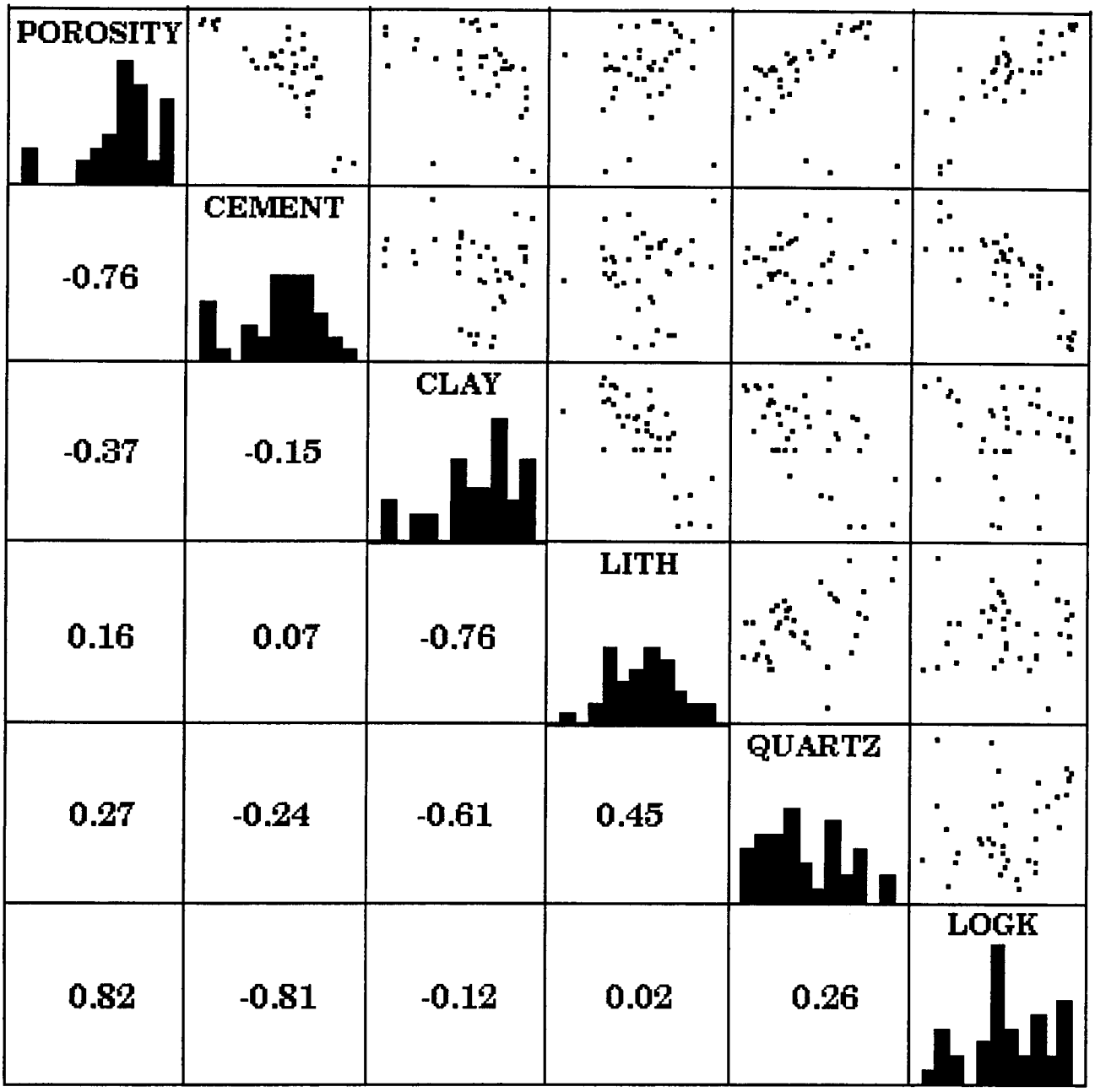
(2) Divide each proportion by the geometric mean of that observation's compositional proportions and take its logarithm

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

(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.. 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.



Picaroon Sandstone scatterplot matrix and gamma correlations of composition 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 $\rho = 0$, 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:

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

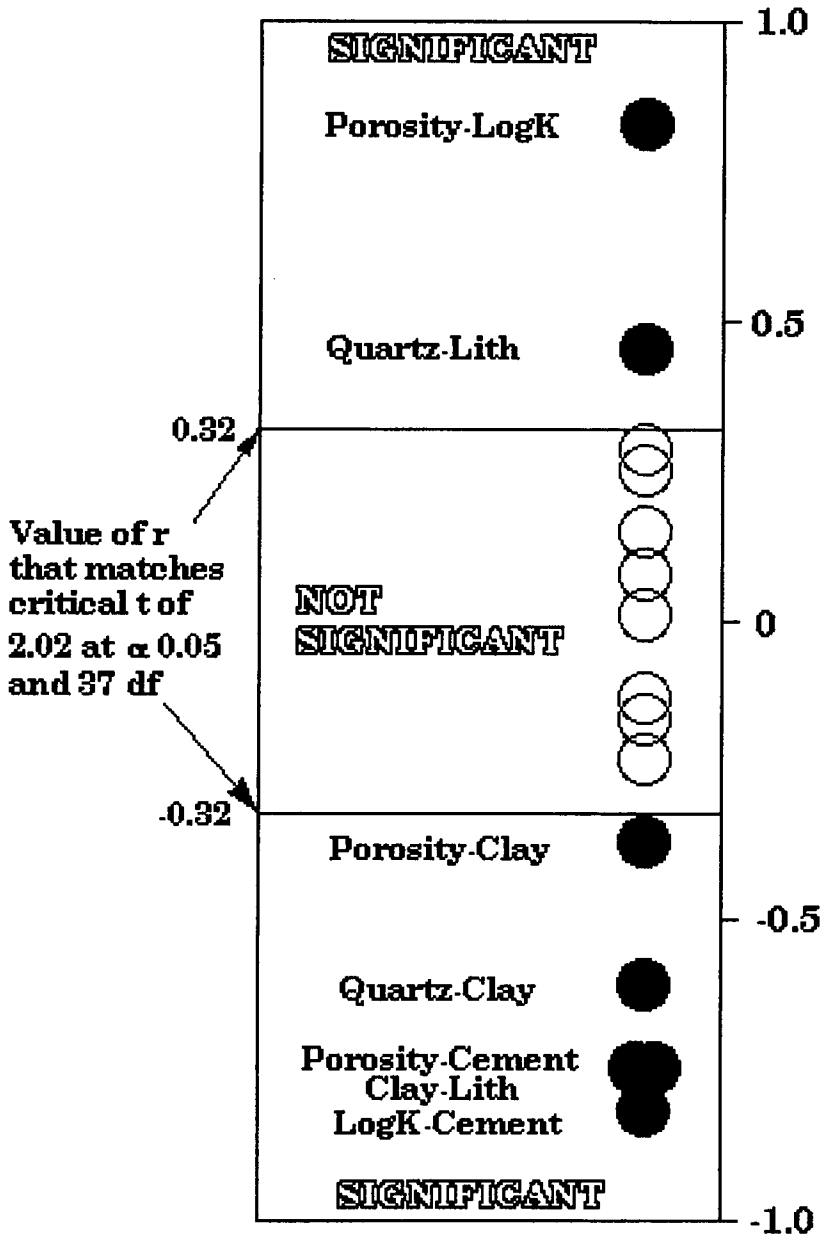
If the calculated value for *t* exceeds the tabulated critical value for a two-tailed test at a selected significance level, then the null hypothesis of no correlation is rejected.

In the case of the Picaroon sandstone data, the sample size, $n = 39$, so that the number of degrees of freedom, $\nu = 37$. 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. 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: labeled closed circles; correlations considered not to be significant: unlabeled 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:

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

POROSITY		SPECIFIC SURFACE AREA		
%	RANKED	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

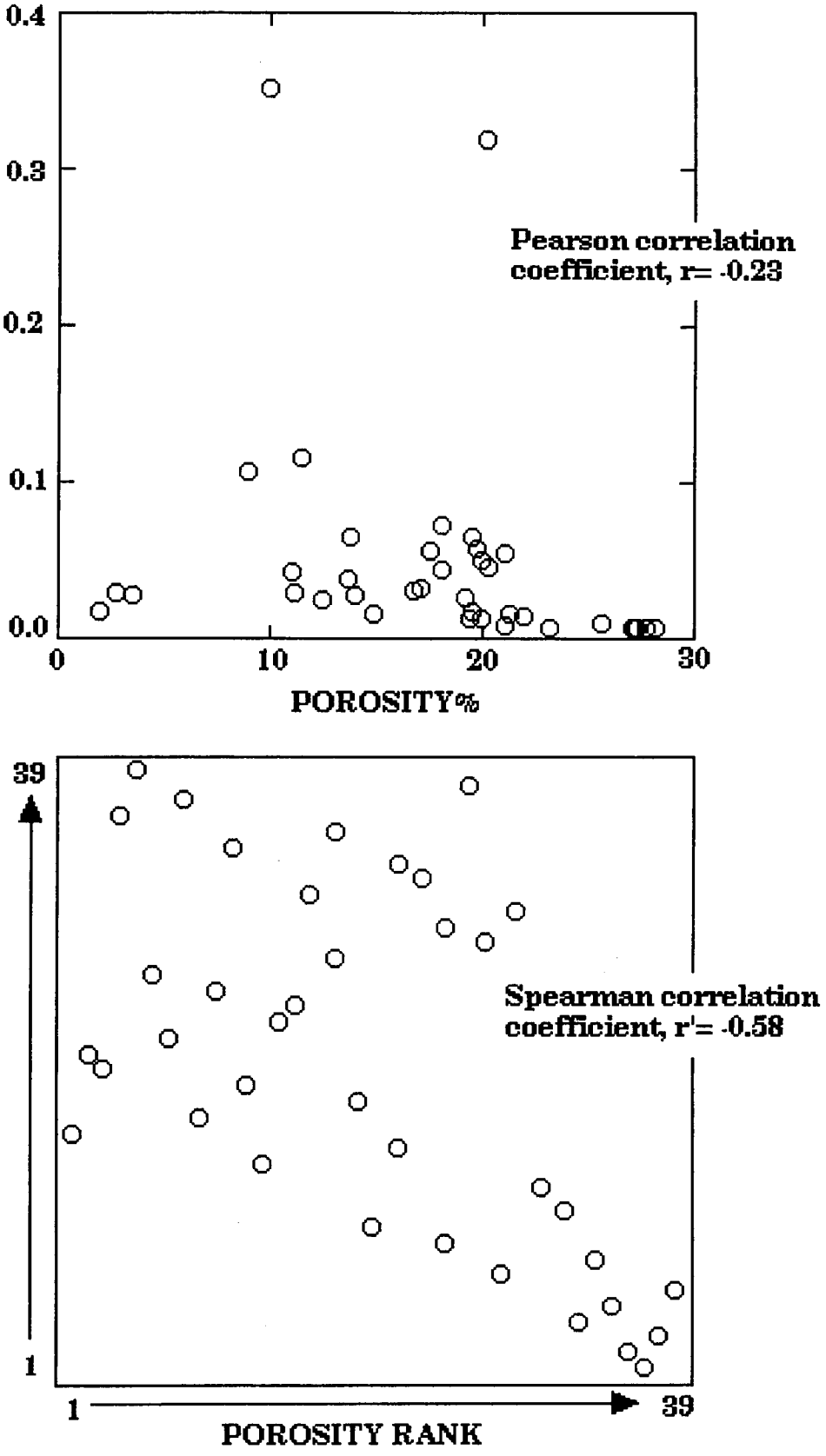
Two-tailed test
critical t = 2.02
@0.05a and 37 df

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

Picaroon Sandstone: Computation of Spearman rank correlation coefficient between porosity and estimated specific surface area.

The values for S are listed in the table . 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. 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: Plots of porosity and specific surface area (a) as raw variables (b) as ranks.

MATRIX ALGEBRA

Matrix theory in its modern usage originated in the work of Cayley and Sylvester in the 1840's. Matrix algebra is the language of multivariate analysis, making possible the succinct description of techniques that handle cumbersome arrays of data and statistics. From an operational point of view, matrices are easily stored and manipulated in a computer. The presentation of data on a spreadsheet as blocks of numbers in cells on a row and column grid takes a matrix form.

Basic matrix terminology

A *matrix* is a rectangular array of *elements* such as:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 4 \\ 7 & 9 \\ 0 & 1 \end{bmatrix}$$

If a matrix has r *rows* and c *columns*, it is said to be of *order* r x c (or known as an r x c matrix). When r=c, the matrix is *square*.

A matrix with only one row is a *row vector* ;

$$\begin{bmatrix} 4 & 9 & 3 \end{bmatrix}$$

a matrix with only one column is a *column vector*.

$$\begin{bmatrix} 5 \\ 1 \\ 9 \end{bmatrix}$$

A matrix with a single row and a single column (i.e. a single number or element) is known as a *scalar*.

$$\begin{bmatrix} 3 \end{bmatrix}$$

In most notations, a matrix is denoted by a single capital letter, conventionally printed in bold type, such as:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

a_{ij} is recognized as the element of **A** that occurs in the *ith* row and the *jth* column. The *leading diagonal* of a square matrix corresponds to the aii elements (on the diagonal reading from upper left to lower right). A *symmetric matrix* is a square matrix such that $a_{ij} = a_{ji}$ for all values of i and j. The matrix is symmetrical about the leading diagonal:

$$\begin{bmatrix} 1 & 2 & 5 \\ 2 & 4 & 3 \\ 5 & 3 & 7 \end{bmatrix}$$

The *trace* of a square matrix is the sum of its leading diagonal elements (12 in the matrix above). A *diagonal matrix* has leading diagonal elements which are non-zero values and zero-valued off-diagonal elements. A diagonal matrix is often denoted as _

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

An important diagonal matrix is the *identity matrix* whose diagonal elements are all one. The identity matrix is denoted as **I**.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The transpose of matrix D is written D" or D^t and is a matrix whose rows are the columns of D (or equivalently, whose columns are the rows of D). Then if:

$$D = \begin{bmatrix} 3 & 7 \\ 1 & 9 \\ 4 & 2 \end{bmatrix} \qquad \text{then} \qquad D^t = \begin{bmatrix} 3 & 1 & 4 \\ 7 & 9 & 2 \end{bmatrix}$$

Addition of matrices

Only matrices of the same order may be added. The sum of the matrices **A** and **B** is the sum of their corresponding elements:

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \\ a_{31} + b_{31} & a_{32} + b_{32} & a_{33} + b_{33} \end{bmatrix}$$

and **A + B = B + A**

Subtraction of matrices operates in the same way, with subtraction rather than addition of individual elements.

Multiplication of matrices

When a matrix is multiplied by a scalar, each element of the matrix is transformed by the scalar:

$$kA = \begin{bmatrix} ka_{11} & ka_{12} \\ ka_{21} & ka_{22} \end{bmatrix}$$

Multiplication of two matrices is more complex and is only possible when the number of columns in the first matrix is the same as the number of rows in the second matrix.

If **C** = **AB**, then each element c_{ij} in the **C** matrix is the sum of the products obtained by multiplying the ith row of **A** by the jth row of **B** or: $c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$ where m is the number of columns in A and rows of B. For example,

if $A = \begin{bmatrix} 3 & 1 & 2 \\ 4 & 0 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 3 & 0 \end{bmatrix}$ then $C = A.B = \begin{bmatrix} 10 & 7 \\ 10 & 8 \end{bmatrix}$

where
$c_{11} = 3x1 + 1x1 + 2x3$
$c_{12} = 3x2 + 1x1 + 2x0$
$c_{21} = 4x1 + 0x1 + 2x3$
$c_{22} = 4x2 + 0x1 + 2x0$

The order of notation in a product expression is important because $AB \neq BA$

B.A in the example above is $\begin{bmatrix} 11 & 1 & 6 \\ 7 & 1 & 4 \\ 9 & 3 & 6 \end{bmatrix}$

In many cases, a change in the order of matrices to be multiplied may result in a situation where the product does not exist because of the mismatch between the number of prefactor matrix columns and postfactor matrix rows.

The minor product moment = $\mathbf{A^tA}$ and the major product moment = $\mathbf{AA^t}$ and consists of the sums of squares and crossproducts of the columns and rows of the matrix **A**, respectively.

The inverse matrix

Division of one matrix by another in a manner similar to a scalar operation is not directly possible. However, the operation may be made through multiplication by an *inverse matrix*.
Now if, **AB** = **C**
then the value of the matrix **B** is conceptually equivalent to the division of **C** by **A**
However, if a matrix A-1 can be found such that :

$$\mathbf{AA^{-1}} = \mathbf{I} \text{ and } \mathbf{A^{-1}A} = \mathbf{I}$$

then by multiplication:

$$\mathbf{A^{-1}AB} = \mathbf{A^{-1}C}$$
$$\mathbf{IB} = \mathbf{A^{-1}C}$$
$$\mathbf{B} = \mathbf{A^{-1}C}$$

$\mathbf{A^{-1}}$ is known as the inverse matrix of **A** and is defined by the property that a matrix multiplied by its inverse yields an identity matrix, The inverse is only defined for square matrices and does not exist for all of these. If a square matrix does not have an inverse, it is called a *singular matrix*.

The widest use of the inverse matrix in practical applications is in simultaneous equations. For example, the following equations:

$$\begin{aligned} 3x + 0y + 2z &= 19 \\ x + 4y + 2z &= 29 \\ 2x + y + 5z &= 35 \end{aligned}$$

may be written in matrix notation as: $\mathbf{AX} = \mathbf{C}$ which in expanded form is:

$$\begin{bmatrix} 3 & 0 & 2 \\ 1 & 4 & 2 \\ 2 & 1 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 10 \\ 29 \\ 35 \end{bmatrix}$$

Solution of the unknowns x, y, and z in the vector $\mathbf{X}$ is given by $\mathbf{X} = \mathbf{A}^{-1}\mathbf{C}$

The elements of the inverse matrix $\mathbf{A}^{-1}$ are: $\begin{bmatrix} 9/20 & 1/20 & -1/5 \\ -1/40 & 11/40 & -1/10 \\ -7/40 & -3/40 & 3/10 \end{bmatrix}$

as solved by the iterative Gauss-Jordan method. When premultiplying the vector $\mathbf{C}$ by the inverse matrix $\mathbf{A}^{-1}$ the resultant vector $\mathbf{X}$ gives the values for x, y, and z.

$$X = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$$

EXCEL matrix algebra functions

The EXCEL matrix functions are:

MDETERM()	returns the determinant of a matrix
MINVERSE()	produces the inverse of a matrix
MMULT()	multiplies two matrices
TRANSPOSE()	transposes the rows and columns of a matrix

AN APPLICATION OF MATRIX ALGEBRA TO POROSITY DETERMINATION FROM LOGS OF COMPLEX RESERVOIRS

Solving the inverse problem was one of the earliest applications of computer processing to the estimation of true volumetric porosity in formations composed of a mixture of minerals. The sonic, density, or neutron tools each give measurements that can be converted to porosity when calibrated to the mineral of the reservoir rock. However, when the rock is composed of several minerals whose proportions vary through the reservoir section, then two or more porosity logs should be run to discriminate lithology effects from porosity.

As described by Savre (1963), the method proved particularly effective in the Permian carbonates of West Texas as a means of improving the log estimates of porosity calculated by extant methods. Most commonly, porosities had been evaluated from neutron logs, but the values often proved to be unduly optimistic in zones with significant gypsum contents. Hydrogen within the water of crystallization of gypsum results in an apparent porosity in excess of 50% for pure gypsum, so that even moderate amounts cause large porosity errors. The grain density and transit time of gypsum are also markedly dissimilar from other common matrix minerals, so that the substitution of an alternative porosity log is not particularly helpful. Finally, the use of a conventional crossplot is difficult because the lithologies are a mixture of three minerals: dolomite, gypsum, and anhydrite. The total system actually consists of four components when pore fluid is considered, and is difficult to represent graphically in a useful manner.

An effective log-analysis solution to this problem is one that computes the contents of gypsum and the other minerals and corrects apparent porosities recorded by the logs to true volumetric porosities. The inverse solution of a log-response equation set achieves this objective very neatly by computing mineral proportions and true porosities simultaneously. Because the unknowns are the fractions of four components, four equations are required for a unique solution. For this application, the log-response equations are:

Neutron: $$\Phi_n = n_g \cdot G + n_a \cdot A + n_d \cdot D + n_f \cdot \Phi$$

Sonic: $$\Delta t = \Delta t_g \cdot G + \Delta t_a \cdot A + \Delta t_d \cdot D + \Delta t_f \cdot \Phi$$

Density: $$\rho_b = \rho_g \cdot G + \rho_a \cdot A + \rho_d \cdot D + \rho_f \cdot \Phi$$

and the model is completed by a fourth "unity" equation which reflects the fact that the unknown proportions of gypsum (G), anhydrite (A), dolomite (D), and true fractional porosity (Φ) form a closed system:

Unity: $$1 = G + A + D + \Phi$$

When written as matrices:

$$\begin{bmatrix} n_g & n_a & n_d & n_f \\ \Delta t_g & \Delta t_a & \Delta t_d & \Delta t_f \\ \rho_g & \rho_a & \rho_d & \rho_f \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} G \\ A \\ D \\ \Phi \end{bmatrix} = \begin{bmatrix} \Phi_n \\ \Delta t \\ \rho_b \\ 1 \end{bmatrix}$$

Rewritten in the more compact matrix algebra convention, $\mathbf{CV} = \mathbf{L}$, $\mathbf{C}$ is the matrix of neutron porosities, transit times, and grain densities of gypsum, anhydrite, dolomite, and pore fluid, augmented by a line of unit values; $\mathbf{V}$ is the vector of their unknown proportions in the zone; and $\mathbf{L}$ is a vector of the zone log readings of neutron porosity, transit time, and bulk density, together with a unit value.

The values of the elements in matrix $\mathbf{C}$ are known:

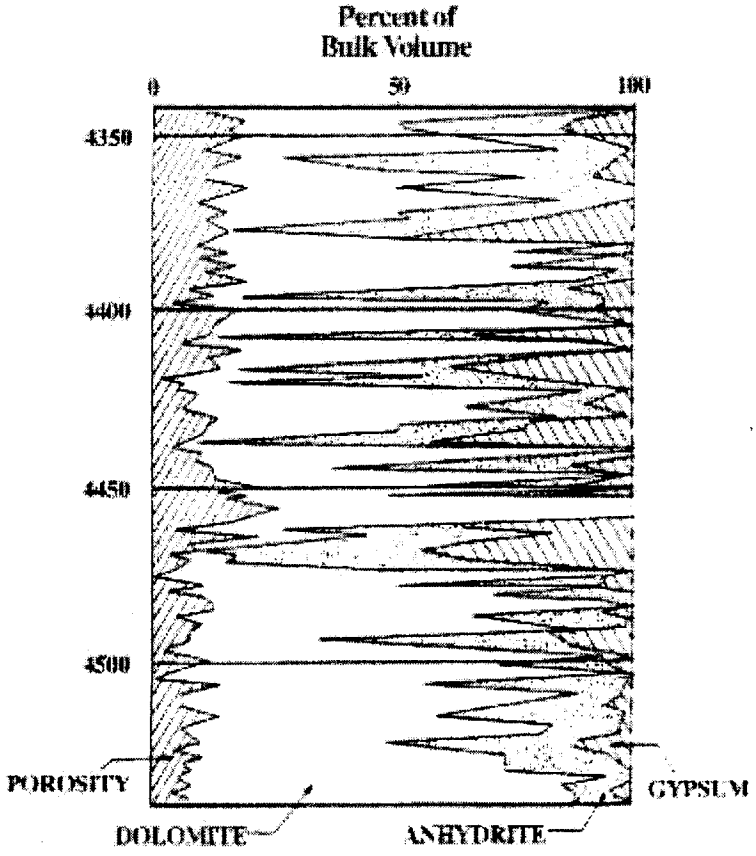
$$\begin{bmatrix} 60 & -2 & 4 & 100 \\ 52 & 50 & 43.5 & 189 \\ 2.35 & 2.98 & 2.85 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

The equation set is fully determined (the number of unknowns is matched by the number of equations) and the solution for the unknown vector, $\mathbf{V}$ is:

$$\mathbf{V} = \mathbf{C}^{-1}\mathbf{L}$$

where $\mathbf{C}^{-1}$ is the inverse of the $\mathbf{C}$ matrix. This procedure can be coded in a computer program or as a spreadsheet solution.

The results can be displayed as a function of depth and the graphical output drafted from one of the earliest computer runs is given below (Alger et al., 1963). Here, profiles of porosity, dolomite, anhydrite, and gypsum are shown from a Permian San Andres Formation section in West Texas. The compositional profile is a numerical transformation of the original log traces which have been processed by the inverse matrix operator.



COMPOSITIONAL ANALYSIS OF A MISSISSIPPI CHAT SECTION FROM LOGS USING A MATRIX ALGEBRA PROCEDURE IN EXCEL

As an example of the EXCEL implementation of this analytical procedure applied to logs, we can analyze the composition of a Mississippian section (see Dataset 5), using a matrix algebra representation. The density-neutron-photoelectric factor log suite provides three logging measurements that can be used to resolve four unknowns in a set of simultaneous equations. The four unknown components are : dolomite (D), chert (Q), calcite (C). and porosity (Φ). The porosity component is the fluid in the pore space of the flushed zone, which is primarily mud filtrate and can be considered to be a mineral called "water".

The equations are:

Neutron porosity: $\Phi_n = \Phi_{nD} \cdot D + \Phi_{nQ} \cdot Q + \Phi_{nC} \cdot C + \Phi_{n\Phi} \cdot \Phi$

Bulk density: $\rho_b = \rho_D \cdot D + \rho_Q \cdot Q + \rho_C \cdot C + \rho_\Phi \cdot \Phi$

Bulk photoelectric factor: $U = U_D \cdot D + U_Q \cdot Q + U_C \cdot C + U_\Phi \cdot \Phi$

Unity equation: $1 = D + Q + C + \Phi$

where the bulk photoelectric factor is the product of the photoelectric factor and the bulk density, $U=Pe^{*}\rho_b$

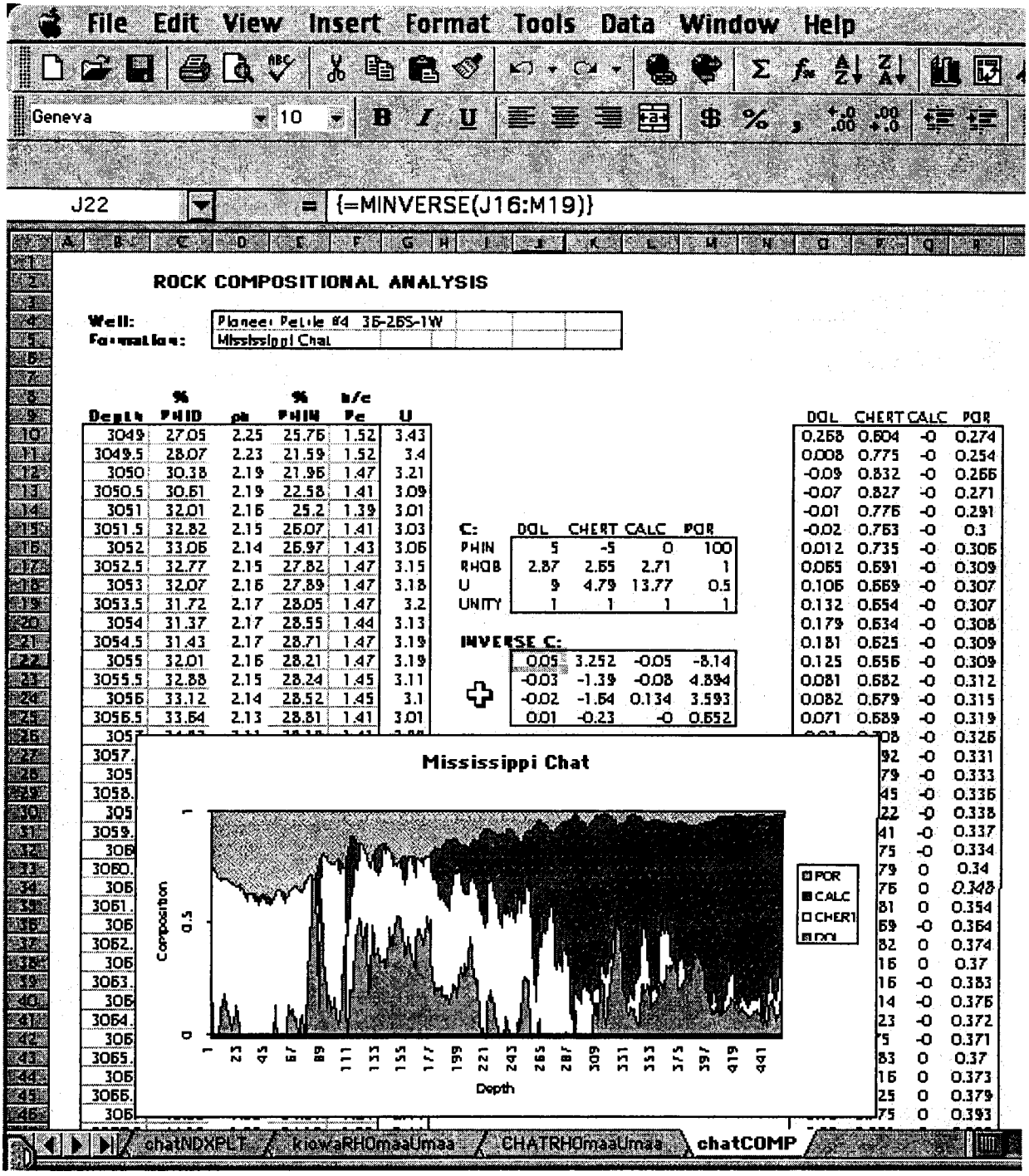
When the values for the log coefficients of the endmembers are entered in the matrix C, together with the unit weights from the unit equation, then C is:

	DOL	CHERT	CALC	POR
PHIN	5	-5	0	100
RHOB	2.87	2.65	2.71	1
U	9	4.79	13.77	0.5
UNITY	1	1	1	1

Using the EXCEL function MINVERSE() applied to these values, the result is:

0.0491	3.2525	-0.0491	-8.1385
-0.0347	-1.3850	-0.0828	4.8935
-0.0202	-1.6356	0.1336	3.5934
0.0058	-0.2319	-0.0017	0.6516

A compositional solution is calculated when the logs are premultiplied by this inverse matrix, which can be graphed by EXCEL as shown on the following page.



The graphic composition plot of the Mississippian section generated by the spreadsheet program shows clearly the microporous spiculitic chert at the top underlain successively by cherty dolomite, chert limestone, and dolomitic limestone units.

COMPUTATION OF AN ARRAY OF PEARSON PRODUCT-MOMENT CORRELATIONS USING MATRIX ALGEBRA

The equation for the correlation coefficient given earlier in this manual was described in terms of conventional scalar algebra:

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

where the sample estimate of the correlation, r, between the two variables of x and y is the covariance of x and y, divided by the product of the standard deviations of x and y. This is perfectly adequate for computing a single correlation, but is inefficient when calculating all possible correlations between a large number of variables. The procedure can be coded as a simple matrix algebra algorithm.

If correlations between m variables are wanted, then an m-by-m correlation matrix may be computed with the following steps:
First,
The matrix **X** is an n-by-m array of the raw data where the n rows are the observations for the m variables of the columns.
U is a 1-by-n unit vector (a row vector of n elements, each with a value of one)
Then,
The matrix **C** is defined as the matrix of covariances between the m variables

i.e. $\mathbf{C} = [\mathrm{cov}(X_j X_k)]$

and so,

$$\mathbf{C} = [\mathbf{X'X} - (\mathbf{UX})'(\mathbf{UX})/n]/(n-1)$$

This concise matrix algorithm for all covariances in the array can be understood from examination of the scalar algebra computation of a single covariance between variable X_j and variable X_k

$$\mathrm{cov}(X_j . X_k) = \frac{1}{(n-1)} \sum_{i}^{n} (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

Rewritten, this becomes:

$$\mathrm{cov}(X_j . X_k) = \frac{1}{(n-1)} \left(\sum X_{ij} X_{ik} - \frac{1}{n} \sum X_{ij} \sum X_{ik} \right)$$

This equation is for one element of the **C** matrix.

The components of the matrix algorithm are:
X'X : premultiplying the data matrix **X** by its transpose gives an m-by-m matrix of the sums of raw squares and crossproducts $[\Sigma X_{ij} X_{ik}]$
(UX)'(UX)/n: Premultiplying the data matrix **X** by the row unit vector **U** gives a 1-by-m row vector of sample totals. Then, premultiplying this sample totals vector by its transpose and dividing by n gives an m-by-m matrix with elements of $[(1/n)\Sigma X_{ij} \Sigma X_{ik}]$. Subtracting **(UX)'(UX)/n** from **X'X** gives an m-by-m matrix of the sums of squares and crossproducts of deviations from the variable means. Multiplication of this matrix by the scalar of (n-1) results in the covariance matrix **C**.

Moving from the covariance matrix, **C** to the correlation coefficient matrix, **R** requires each element to be divided by the product of the matching standard deviations. The matrix operation involves these steps:

The matrix **D** is the diagonal form of the covariance matrix, in other words a matrix with zeroes on the off-diagonal elements and the variances of the variables retained on the leading diagonal. The matrix $\mathbf{D}^{1/2}$ is a diagonal matrix with standard deviations on the leading diagonal. The inverse of this diagonal matrix $\mathbf{D}^{-1/2}$ is simply a diagonal matrix with the reciprocals of the elements of $\mathbf{D}^{1/2}$

Then, the correlation matrix, **R** is solved by: $R = D^{-1/2} C D^{-1/2}$

Matrix algorithms will be used to explain the computations used in regression analysis and it will be seen that matrix representation shows that a wide variety of regression models are members of a simple hierarchy.

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:

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

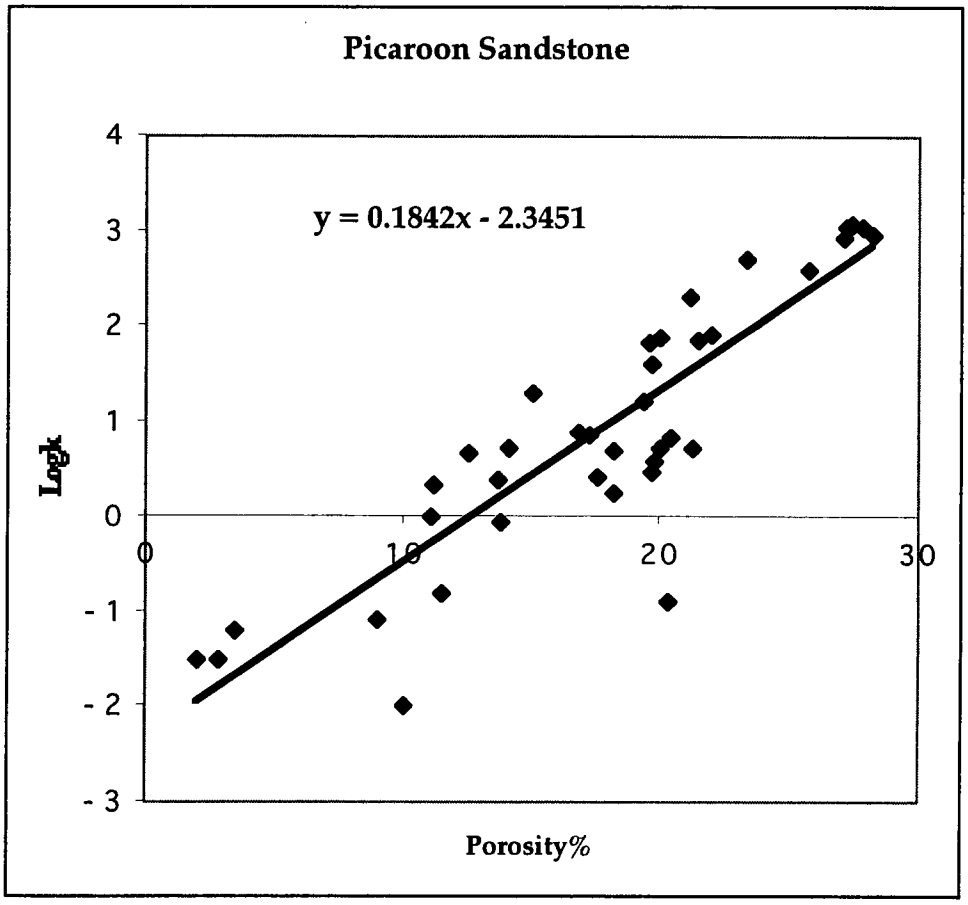
where X is called the independent variable (which is assumed to have no error) and Y is the dependent variable with an associated random error, e. The equation of the line itself is:

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

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).

There are a number of options within EXCEL to fit lines to data and examine their associated statistics. First, we will look at the graphic result of crossplotting porosity percent as the X-axis against the logarithm of permeability as the Y-axis, adding a trendline and selecting the option to add the equation of the line to the plot. The result is shown on the next page.

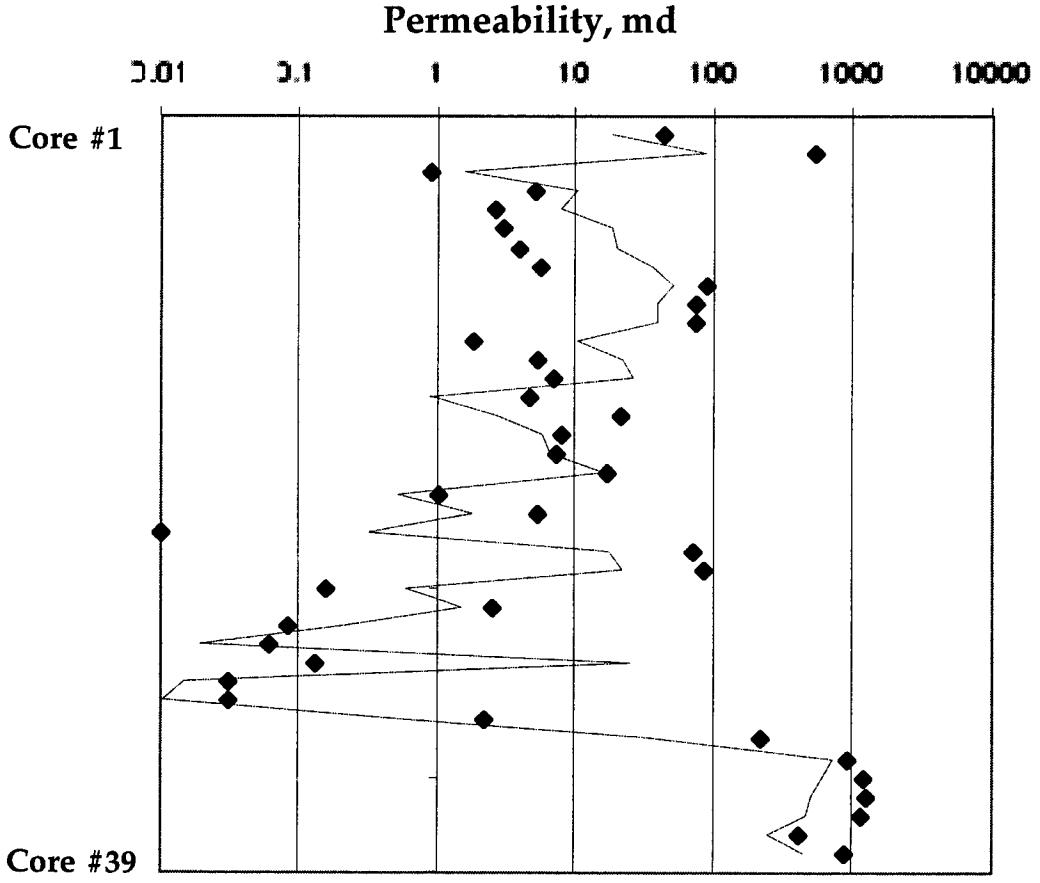


The equation of the fitted line is: $\log k = -2.3451 + 0.1842\Phi$
If it is used to predict permeability based on a percentage porosity then the value in millidarcies is: $10^{\log k}$
The coefficients of this line are -2,3451 which is the estimate of the intercept, a_0 and 0.1842 which is the estimate of the slope, a_1 .
The corresponding parameters are α_0 and α_1 . We can estimate the standard error of both of these quantities using formulae described later in the ANOVA regression analysis section. However, both of these standard errors are given by the EXCEL function LINEST(x,y,TRUE,TRUE) where x is the array of the independent variable, y is the array of the dependent variable, the first TRUE asks for a solution for an intercept value (FALSE would fit a line with a zero intercept), the second TRUE asks for standard errors of the intercept and slope to be output. When LINEST is used with the Picaroon Sandstone porosity - log permeability data, the result is:

0.18419872	-2.3451485
0.0163483	0.30893523

where the top two cells give (again) the slope and intercept, and the cells below record their respective standard errors. In each case, the standard error is the standard deviation of each statistic about the population parameter. We can now make statements about the parameters of the intercept and slope. We can say with 95% confidence that the parameter will be located within plus or minus 1.96 standard errors of the calculated statistic. So, the 95% confidence on the slope parameter is 0.184 +/-0.032 (between 0.152 and 0.216) and the intercept parameter is -2.345 +/- 0.606 (between -1.739 and -2.951).

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. However, a comparison between predicted and actual values of permeability on a depth profile has both good and bad features. 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 (see illustration on the next page). It is the optimal choice, because it generates the minimum potential squared error when the prediction is compared with the actual value. 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 already estimates the mean values.

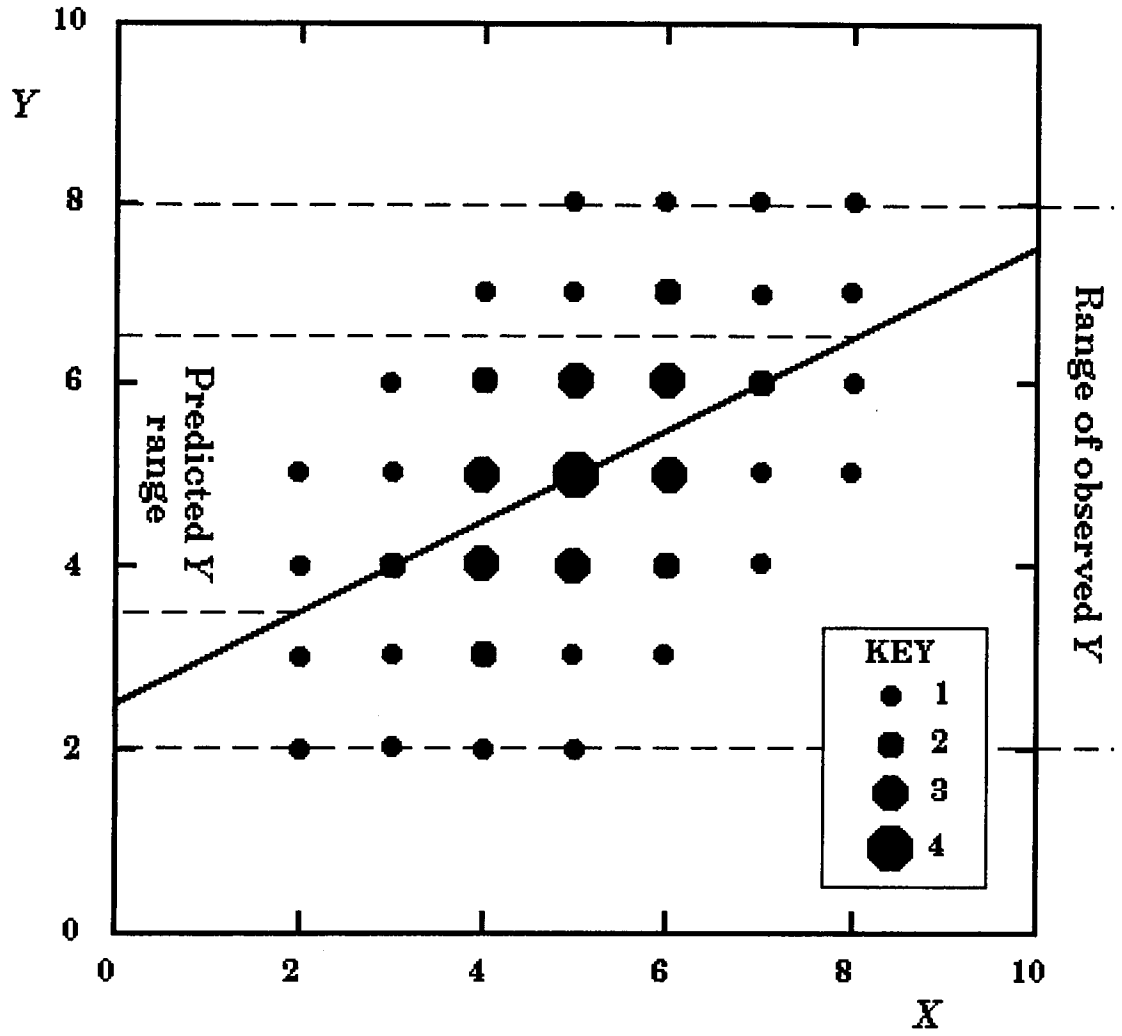
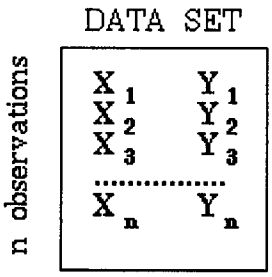


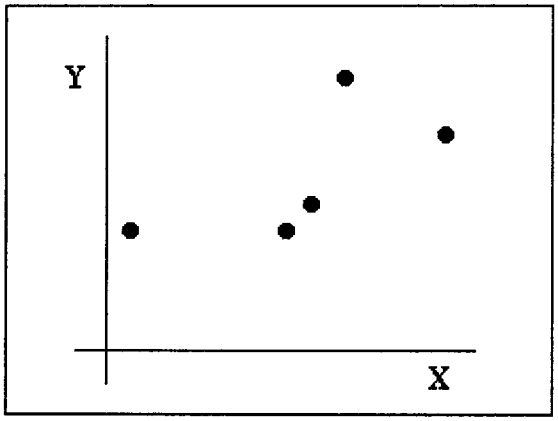
Illustration redrawn from Campbell and Stanley (1966)

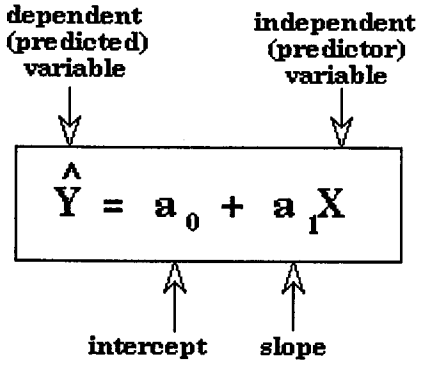
In the illustration above, a regression of Y on X is shown for artificial data and demonstrates that the effects associated with the best prediction of Y on average for any given X value, including the contraction of the prediction range for Y compared with the range of the observed Y values. (The size of the symbols represents the number of points at each location as defined in the key.)

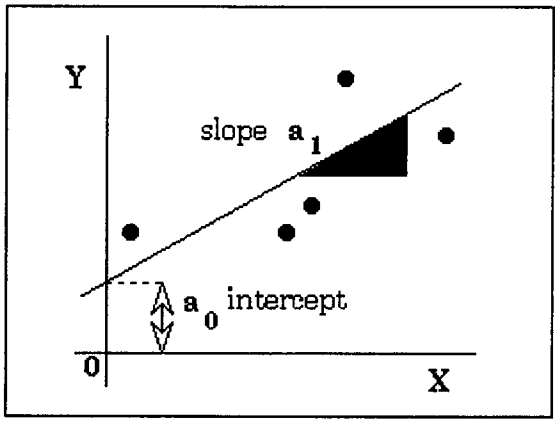
THE GEOMETRY AND CALCULUS SOLUTION OF THE REGRESSION LINE, AND ITS STATISTICAL ASSESSMENT

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 . 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.





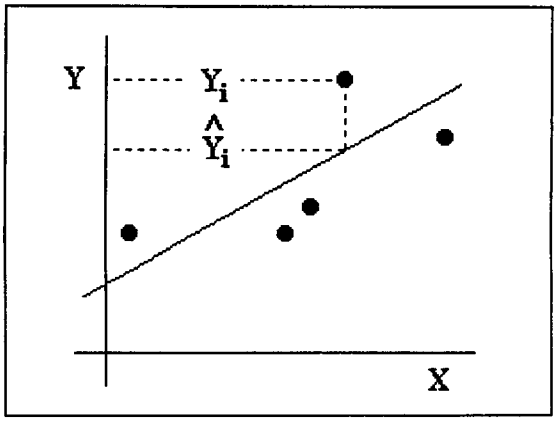




The regression line of Y on X is fitted using the "principle of least squares", which minimizes the sum of the squared deviations of Y from its predicted value, $\hat{Y}$

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

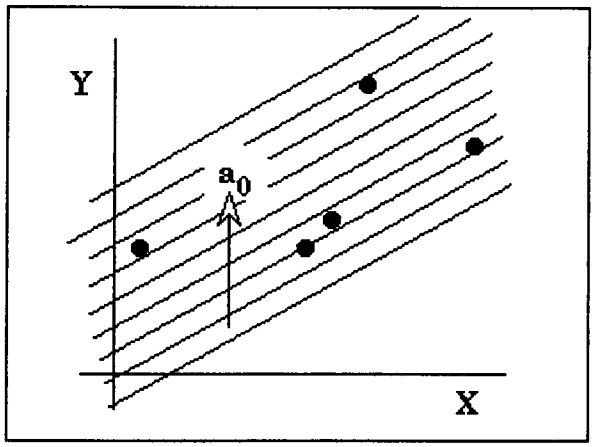
where G is the minimum possible value.

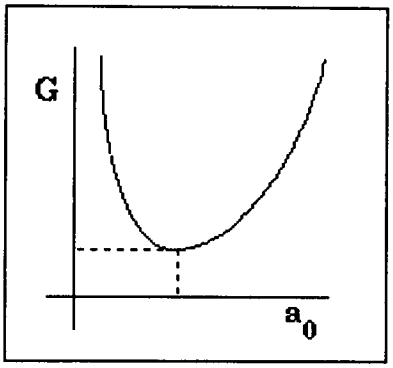


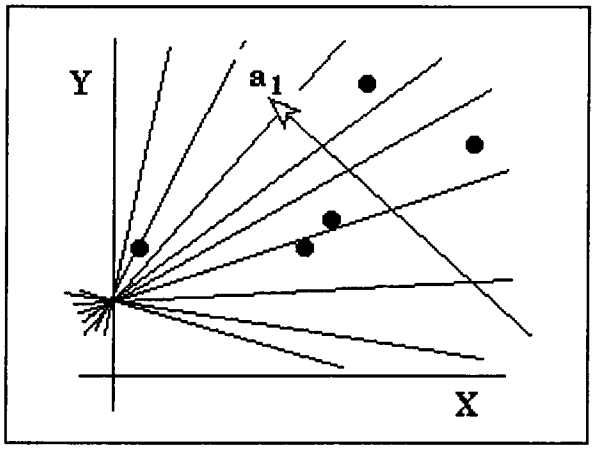
$\sum(Y_i - \hat{Y}_i)^2 = G$ = minimum

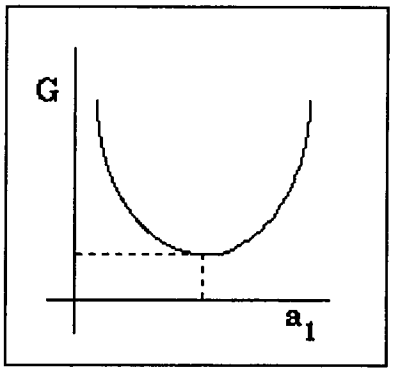
But ... $\hat{Y}_i = a_0 + a_1 X_i$

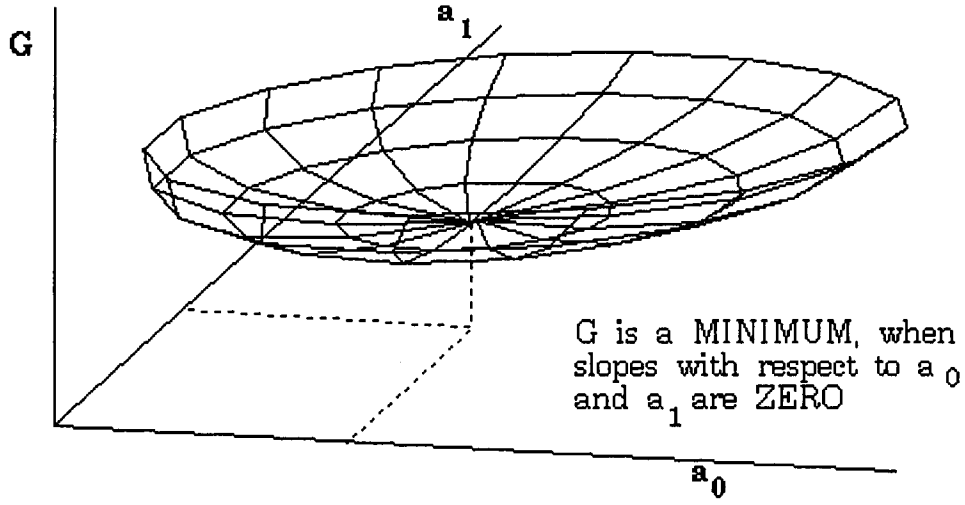
So ... $\sum(Y_i - a_0 - a_1 X_i)^2 = G$ = minimum











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 :

$$na_0 + a_1\sum X_i = \sum Y_i$$
$$a_0\sum X_i + a_1\sum X_i^2 = \sum X_iY_i$$

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

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. 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 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.

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.

We have already seen that we can fit a trend line to a crossplot in EXCEL together with its descriptive equation. A more complete regression analysis can be done by selecting Regression from the Data Analysis add-in on the Tools menu. The output for regression statistics and ANOVA of log permeability on porosity% in the Picaroon Sandstone is shown below:

Regression Statistics	
Multiple R	0.87995426
R Square	0.7743195
Adjusted R Square	0.76822003
Standard Error	0.67435488
Observations	39

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	57.7304454	57.7304454	126.948594	1.6103E-13
Residual	37	16.8259168	0.45475451		
Total	38	74.5563622			

In addition, the regression output lists the coefficients a0 (the intercept) and a1 (the slope) as well as the standard errors of the intercept and slope together with 95% confidence bounds on the parameters:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-2.3451485	0.30893523	-7.5910688	4.6924E-09	-2.9711102	-1.7191869
PHI%	0.18419872	0.0163483	11.2671467	1.6103E-13	0.15107395	0.21732349

These standard errors are also given by the LINEST() function as described earlier together with the explanation of how the confidence range of the parameters is calculated.
The equation to calculate the standard error of the intercept is:

$$s_{a0} = \frac{s\sqrt{\sum x_i^2}}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}}$$

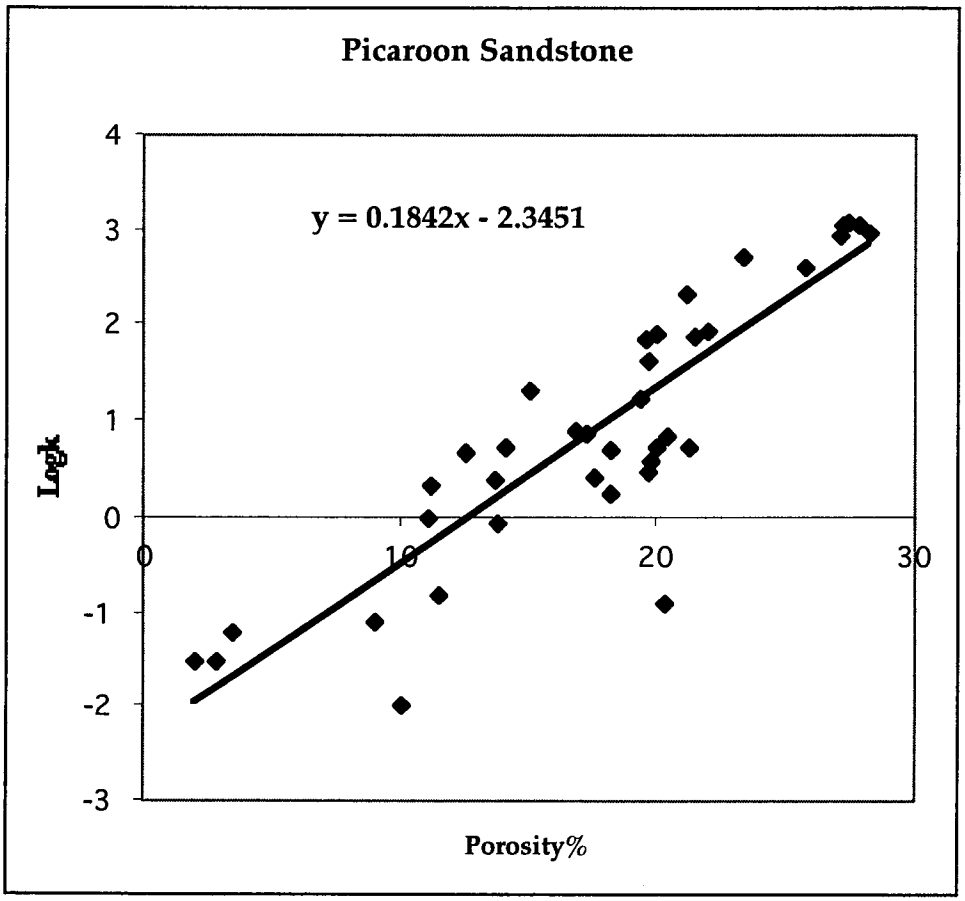
and the standard error of the slope is:

$$s_{a1} = \frac{\sqrt{sn}}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}}$$

where s is the standard deviation of y.

REMOVAL OF OUTLIERS

When fitting a regression trend to observations of an independent variable X to predict a dependent variable Y, the regression model expectation is that residuals of Y (differences between the predicted Y and the observed value of Y) will be normally distributed. "Outliers" are recognized as a few isolated observations that are not well fitted by the trend. They can have a major influence on the trend because they contribute squared deviations to the trend calculation that are considerably larger than more representative points. Consequently their recognition and possible elimination should be considered, but in a conservative approach so that potentially useful information is not lost. As an example, can one or more of the permeability deviations from the prediction of permeability in the Picaroon Sandstone be considered as outliers?

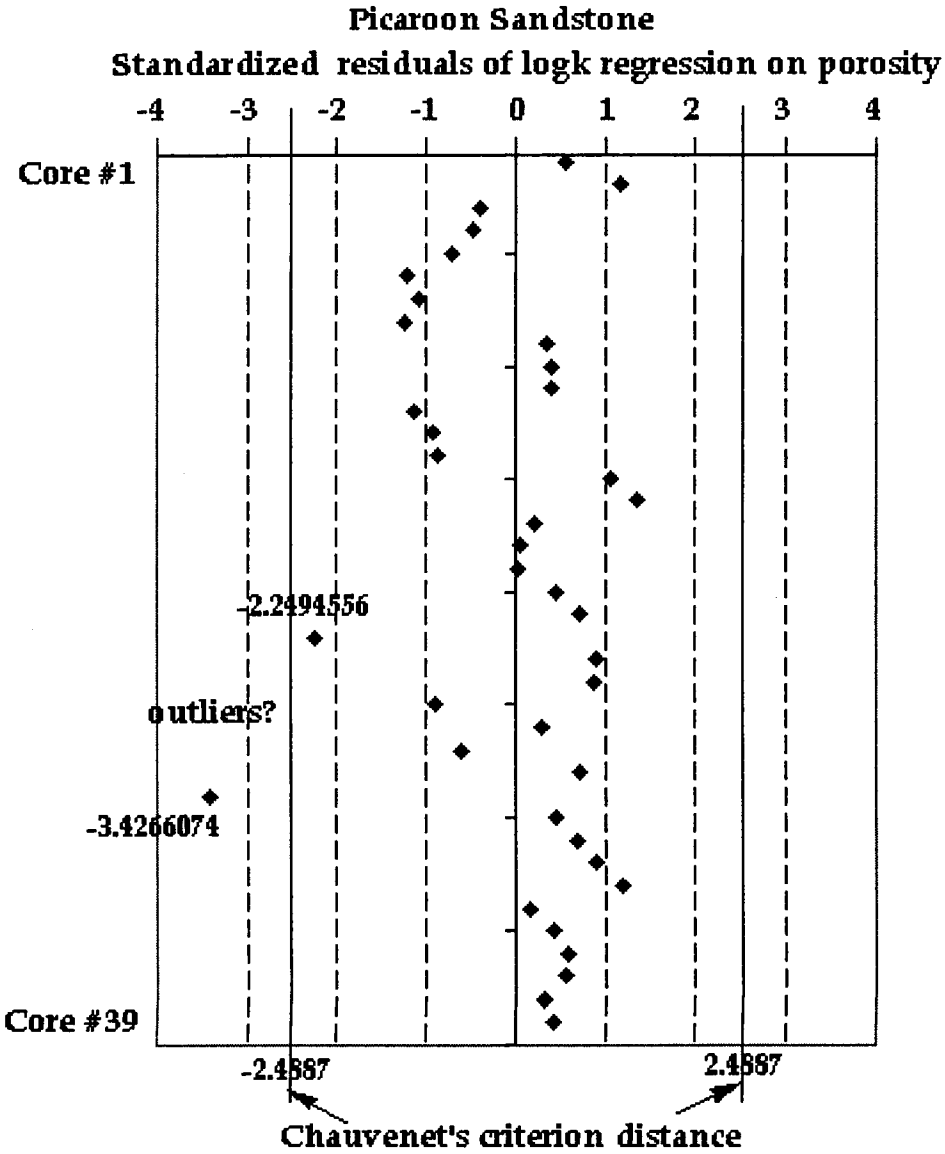


Whether or not an apparently extreme observation is an outlier is best considered by examination of the standardized residuals. The standardized residual is simply the residual divided by the standard deviation:

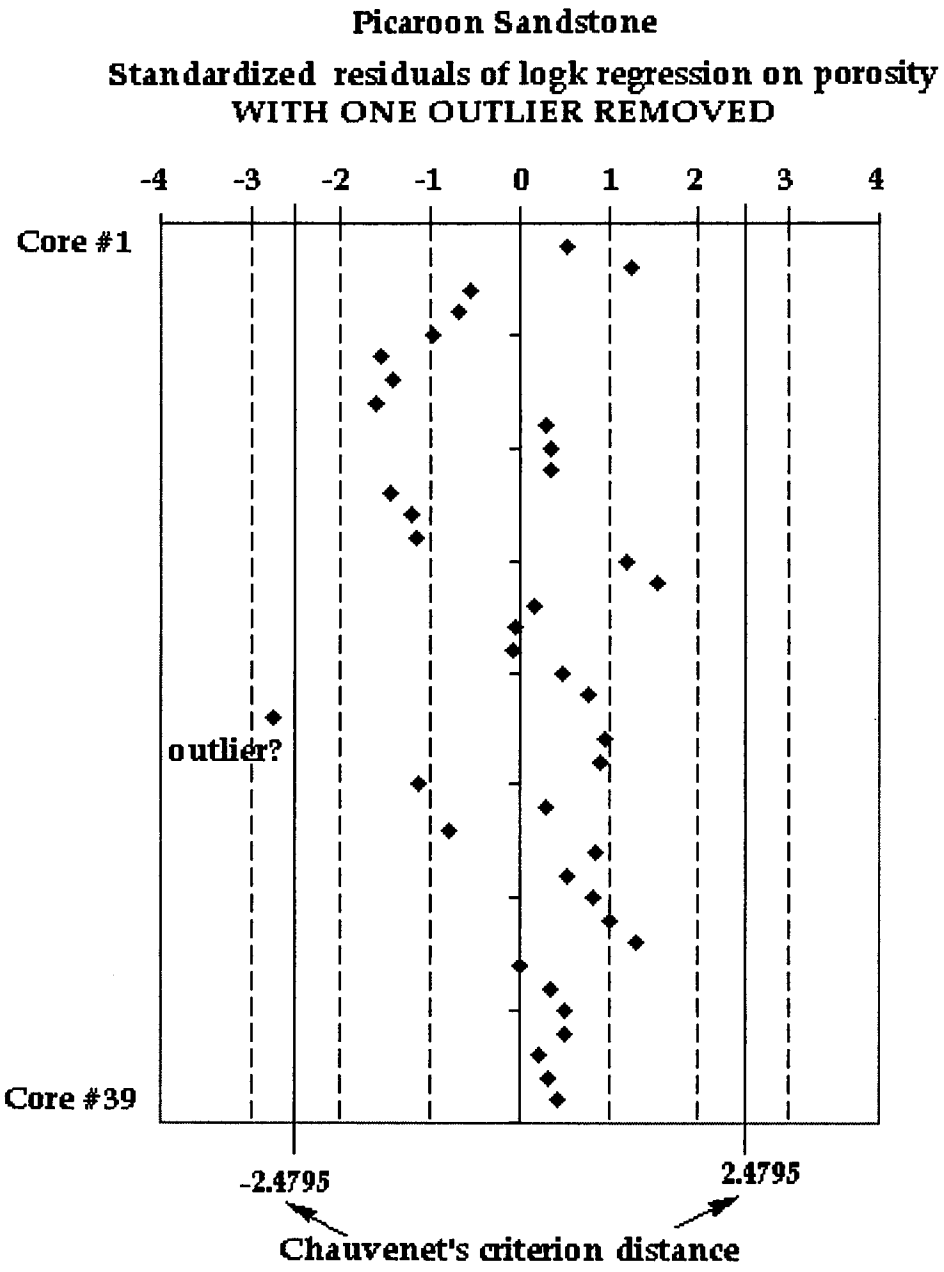
$$Z_r = \frac{Y - \hat{Y}}{s_Y}$$

As a general rule, any observation with standardized residual greater than 2.5 in absolute value should be examined as a possible outlier. Recall that the mean value plus or minus 2.5 standard deviations should contain about 99% of the observations if they are normally distributed (and this is the assumption of the regression model about the residuals). So, an observation with standardized residual greater than 2.5 (or less than -2.5) should occur only 1% of the time. However, the sample size must also be considered. If we have 1000 observations, we would normally expect about ten of these to generate standardized residuals higher than 2,5, as contrasted with a sample of ten observations, even one would be highly anomalous and so an outlier.

Chauvenet's Criterion is a widely used as a means of assessing whether one piece of experimental data — an outlier — from a set of observations, is spurious. Chauvenet's criterion rejects any data points that have less than a 1/(2*n) chance of occurring. There are 39 observations in the Picaroon Sandstone sample, so the critical probability is 0.0128. This probability is for both tails, so that the probability in one tail is 0.0064. The Z standardized residual matched with this (one-tail) probability is 2.4887 and can be obtained from the EXCEL function NORMSINV(p).

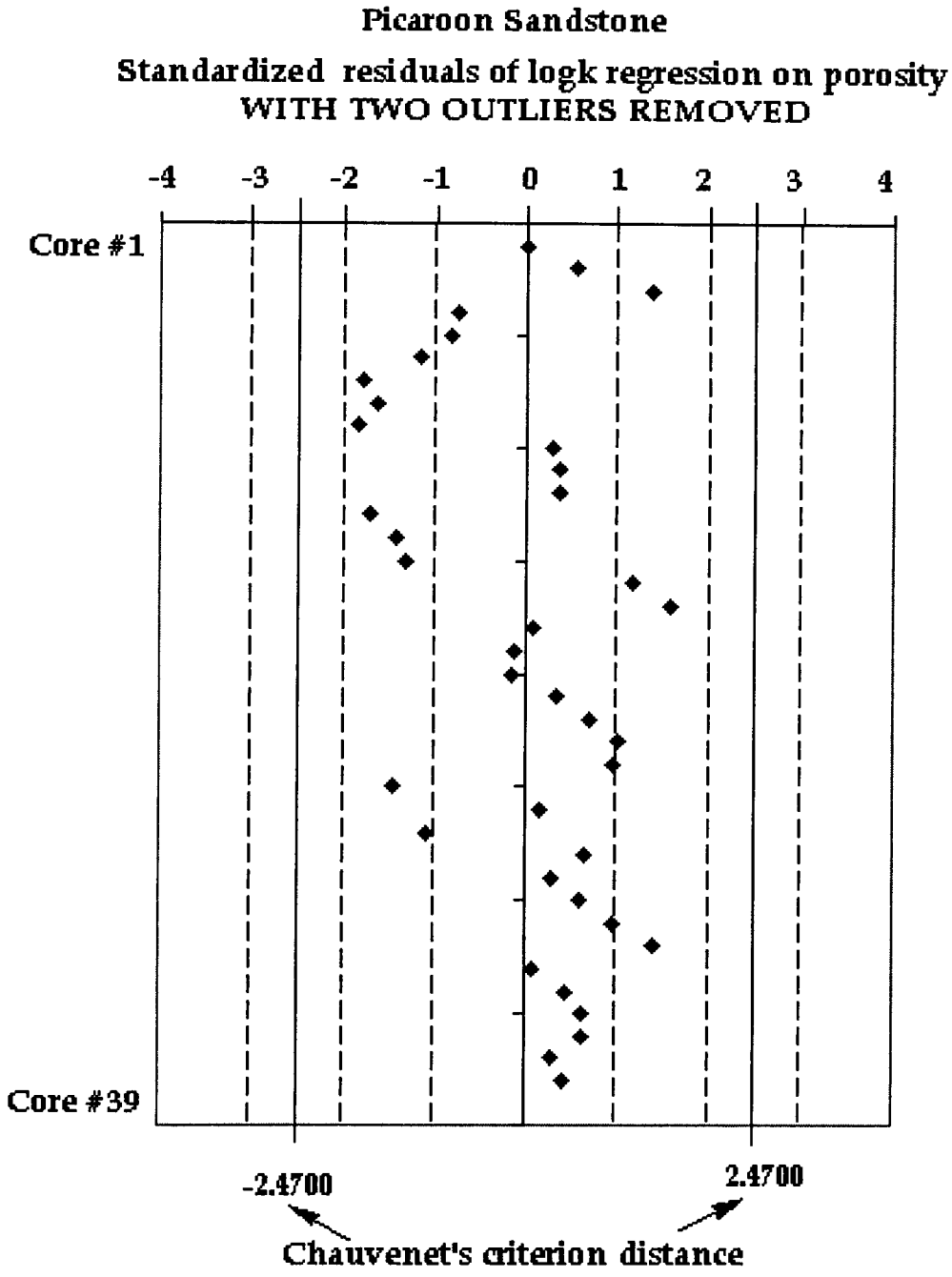


One permeability observation is rejected as an outlier. If we remove it and rerun the regression, the standardized residuals are now:



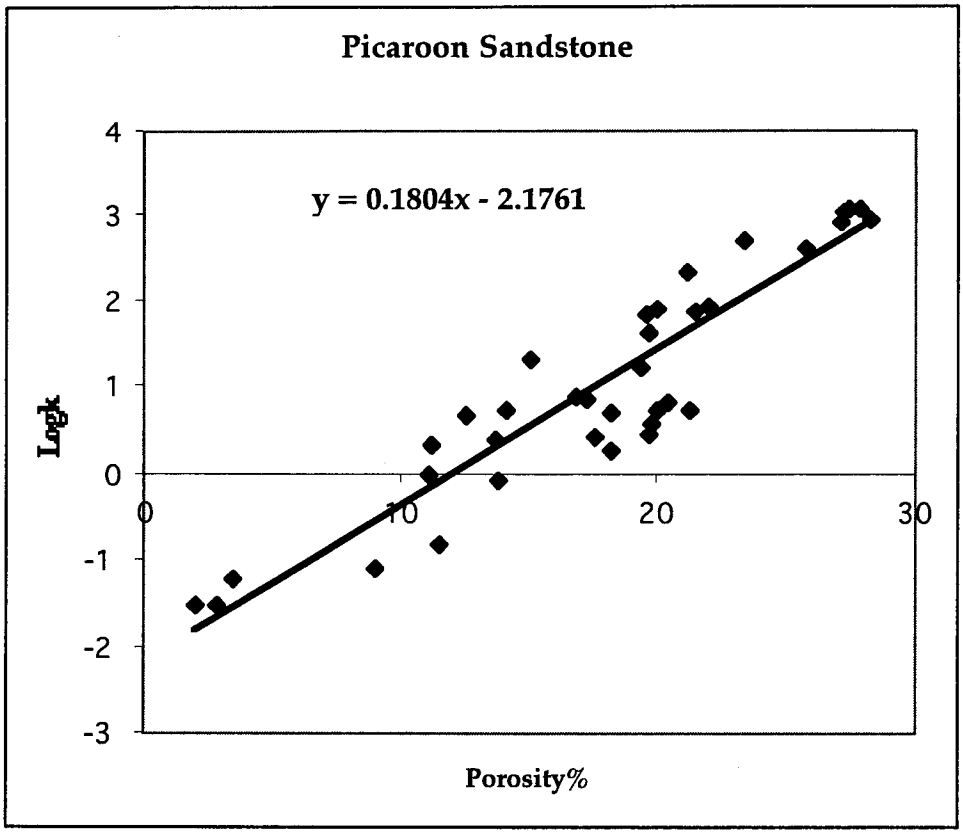
There are two schools of thought on what we do now. The old school says that you are only allowed to apply Chauvenet's criterion once. In this case, the contraction of the standard deviation scale resulting from the removal of the outlier has created a new outlier, but we are not allowed to remove it. The newer school of thought says that we should iterate the process until there are no outliers.

On the next iteration with the removal of the second outlier, the result is:

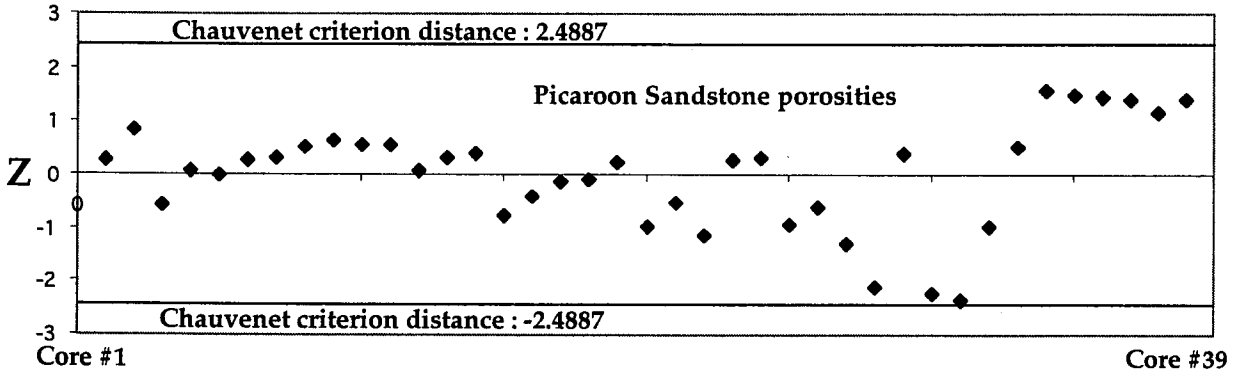


This is the final iteration because all the points lie within Chauvenet's criterion distance. Notice that the distance shrinks on each residual plot because the number in the sample decreased from 39 to 38 (one outlier removed) to 37 (two outliers removed).

The revised linear regression (with two outliers removed is:

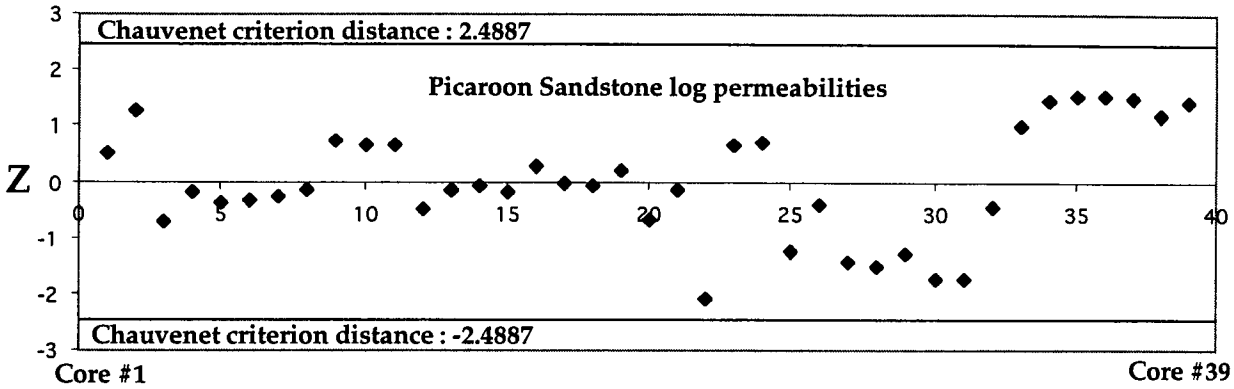


Chauvenet's criterion is also applied to univariate data. For example, we could ask whether the two outliers that were removed were already outliers in the sense of being too far removed from the sample distribution of porosity and permeability that they are likely to be observations from a different population.



None of the core porosities are considered outliers, although three of the low porosity values are close to outlier status

The standardized scores of the logarithmically-scaled permeabilities also lie within Chauvenet's criterion distance calculated for the core sample size of 39 observations.



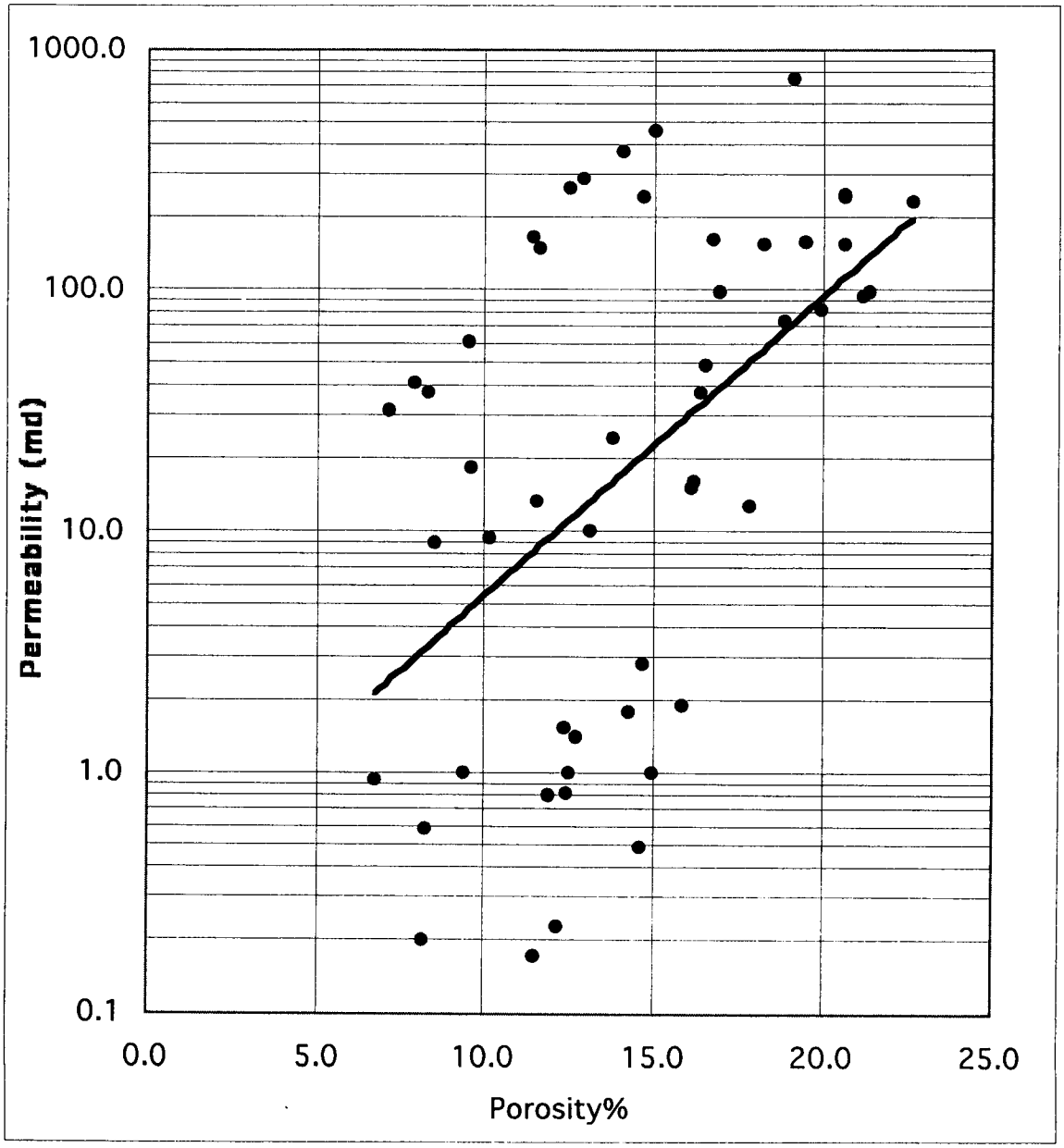
So, the outliers on the regression of log permeability on porosity in the Picaroon Sandstone isolate two core samples in which the permeabilities are anomalously low, given the porosity values.

What do outliers represent?

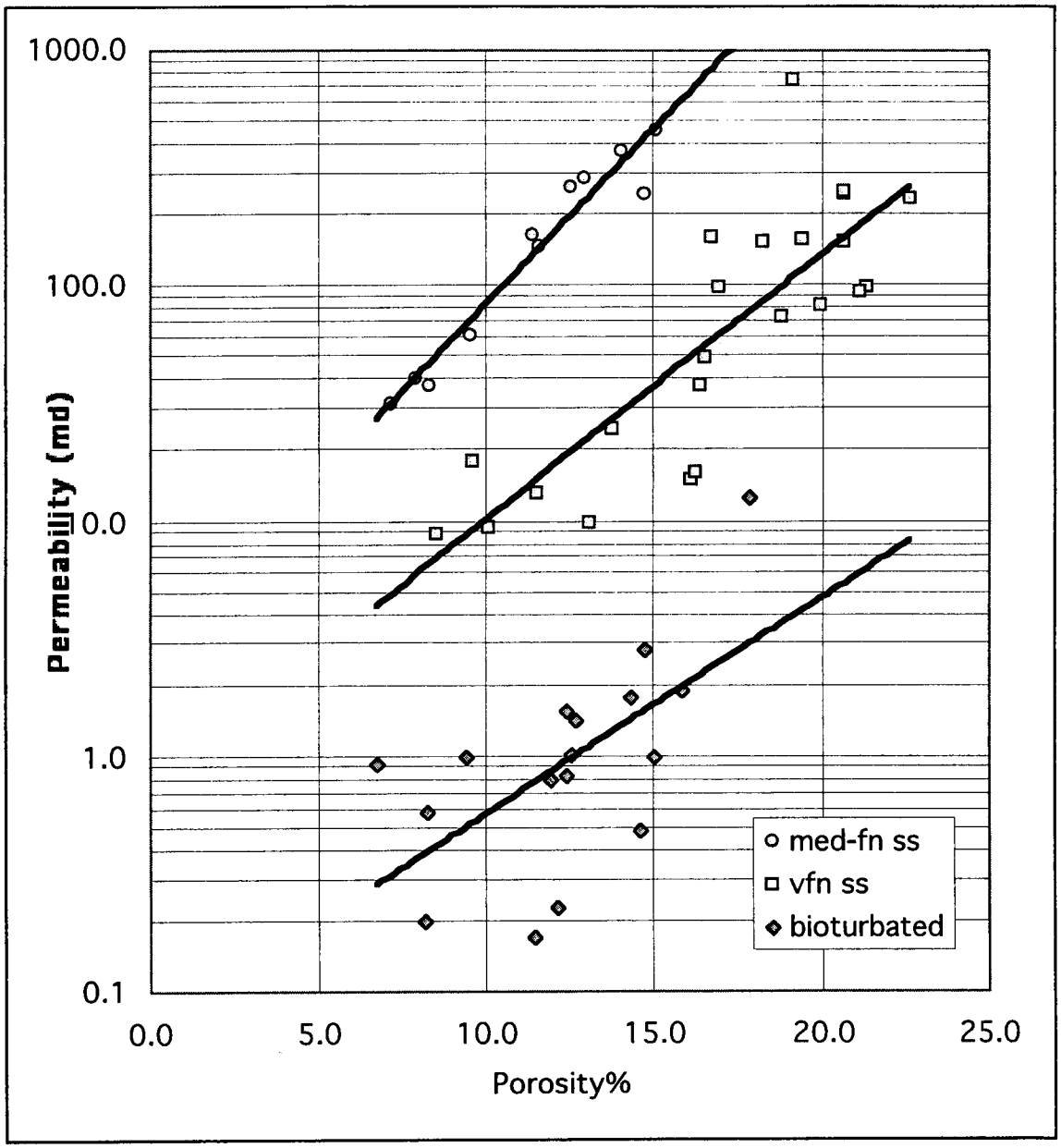
In the case of univariate distribution data, outliers are typically either systematic errors or they may reflect the fact that the population is markedly non-normal in its distribution. When fitting a linear trend to bivariate data, the outliers could again represent mistakes in data recording or other errors of a non-statistical type. Otherwise, the outliers may be caused because there is a functional relationship between the two variables, but it strongly non-linear, or the outliers are sample observations from a different population than that fitted by the trend.

ADDITIONAL CONTROLS ON PERMEABILITY

An example of fitting core measurements of permeability to porosity is shown for the Simpson Sandstone (Middle Ordovician) from core measurements in Kansas reservoirs.



When applied to special cases of homogeneous sandstones, the results may be adequate, but prediction errors are often large in typical sandstones. Notice that the range of predictive error is reduced considerably in the Simpson Sandstone data set if the cores are subdivided by grain-size observed in the core samples. Smaller grain sizes cause greater surface area which decreases permeability; larger grain sizes in rocks of equivalent porosity causes a reduction in surface area and so, increased permeability.



So, permeability is not exclusively determined by pore **volume**, but is also controlled by internal surface area. The interrelationships are contained in the classic Kozeny-Carman equation :

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

which incorporates the specific surface area, *S*. as an additional variable to estimate permeability. 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. Wyllie and Rose (1950) proposed a modification of the Carman-Kozeny equation that substituted irreducible water saturation for the specific surface area term:

$$k = \frac{P\Phi^Q}{S_{wi}^R}$$

The irreducible water saturation term in the modified equation functions as a powerful surrogate variable for specific surface area, and this accounts for the improvement in permeability estimates when incorporated with porosity.

The Wyllie-Rose relationship is a generalized equation that requires the determination of values for the constants P,Q, and R to be calibrated from core measurements. Probably the most widely-used version of this equation is the "Timur equation" **for sandstones**. Timur (1968) developed an equation using regression analysis in which he linked permeability with both porosity and irreducible water saturation (Swi) in sandstones, based on laboratory measurements of core.

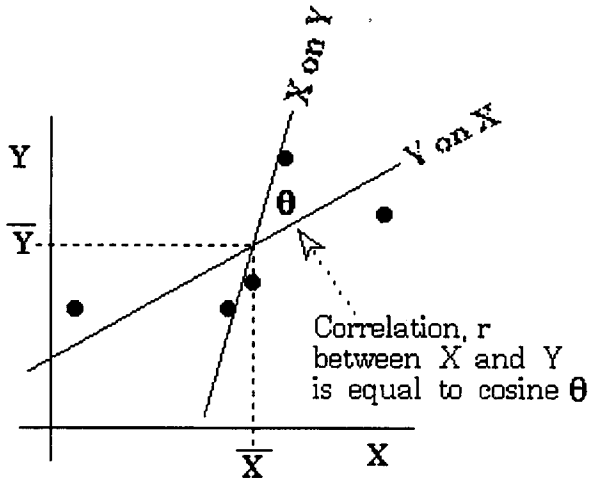
$$k^{0.5} = 100\frac{\Phi^{2.25}}{S_{wi}}$$

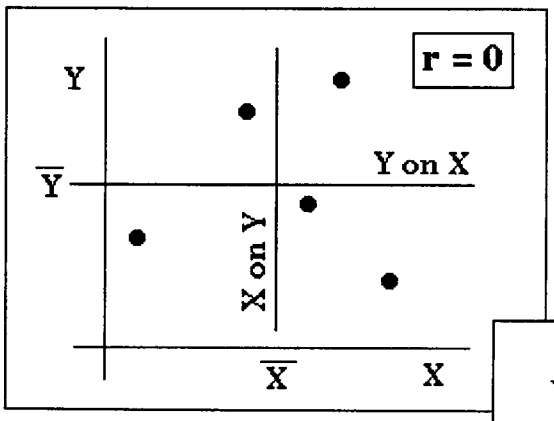
(Both water saturation and porosity values are in fractional units.)
The results showed a considerable improvement in permeability estimation over those based on porosity values alone. However, notice that the use of irreducible water saturation as an input variable restricts the predictions to hydrocarbon reservoir zones.

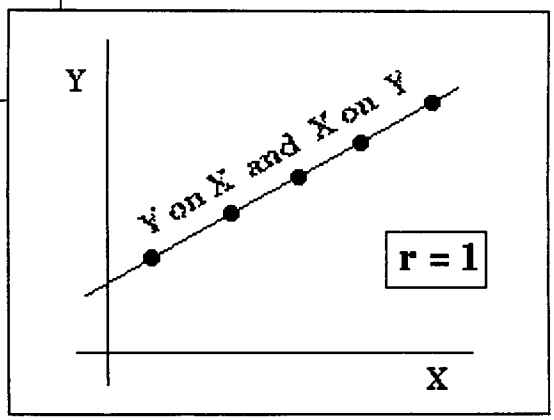
A more advanced logging solution to estimating permeability is the NMR (or MRI) tool which measures pore size distribution as well as porosity through the measurement of nuclear resonance. The computation of permeability from the logging measurement is based on a (continually improved) variant of the Kozeny equation.

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

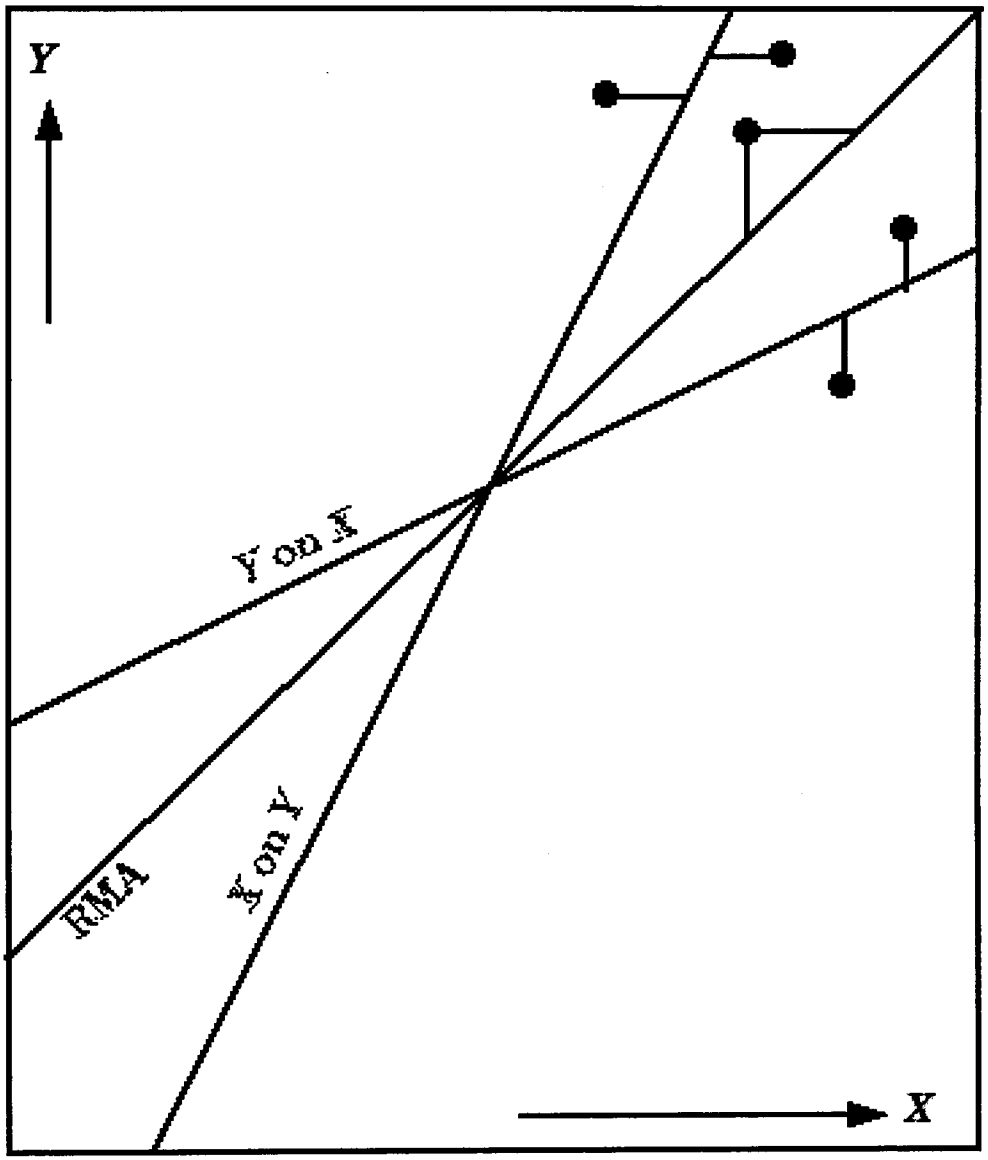
For any set of bivariate data, *x* and *y*, two alternative regression lines may be computed: Y on X or X on Y. One can either predict Y **given** a value of X, or one can predict a value of X **given** a value of Y. The two alternatives attribute all the error to Y (Y on X) or all the error to Y. The regression lines intersect at the coordinates of the bivariate means . 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.



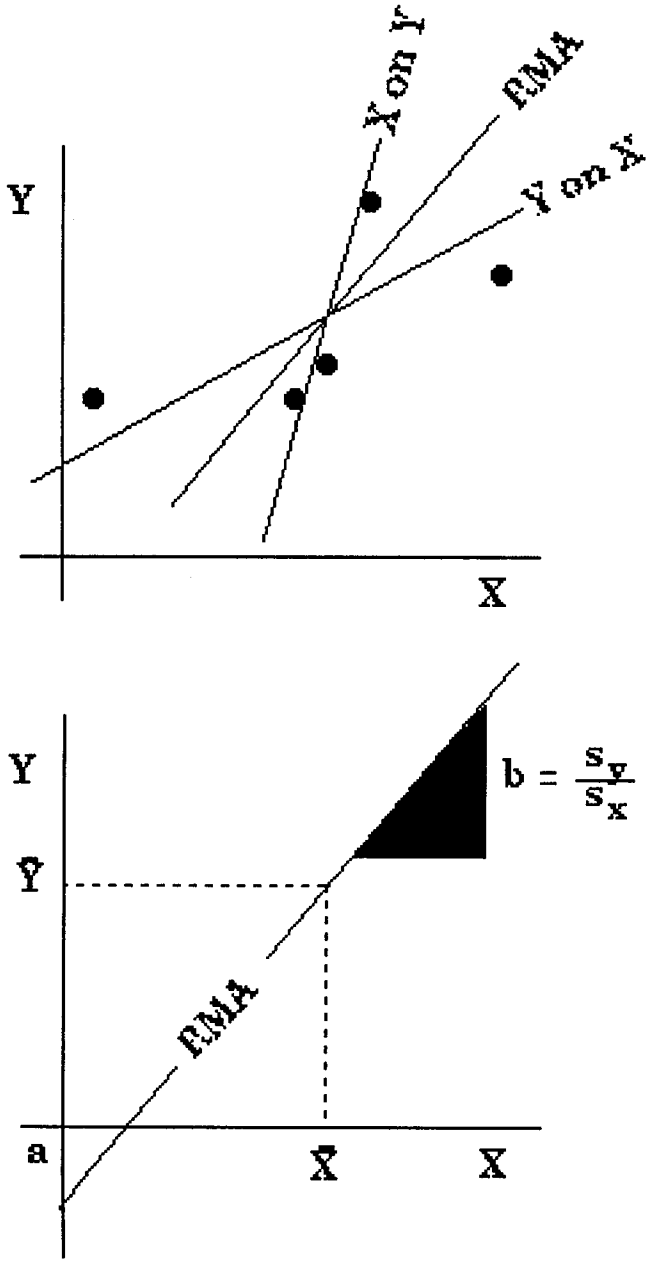




At low correlation coefficients, the divergence between the two lines is large and neither of them appears to be a "best fit" as seen by the human eye. A visually pleasing line would generally be chosen at a position that approximately bisects the two lines, because the human tends to minimize the spread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



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.



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.

ESTIMATION OF POROSITY FROM SONIC LOG TRANSIT TIMES IN THE SADLEROCHIT SANDSTONE OF THE PRUDHOE BAY FIELD, ALASKA

Prudhoe Bay is the largest oil field in North America. There is an estimated 25 billion barrels of oil in Prudhoe Bay trapped within the Sadlerochit Sandstone found 9,000 feet below ground surface. 10 billion barrels of oil have already been produced from the bay, while 13 billion barrels are classified as recoverable with current technology. Although the Sadlerochit Sandstone has been extensively cored, not every well is cored and porosities are estimated from the sonic log using calibration equations developed for sonic logs in wells where core porosity measurements are available.

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 these 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. 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 as will be shown in a later section.

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

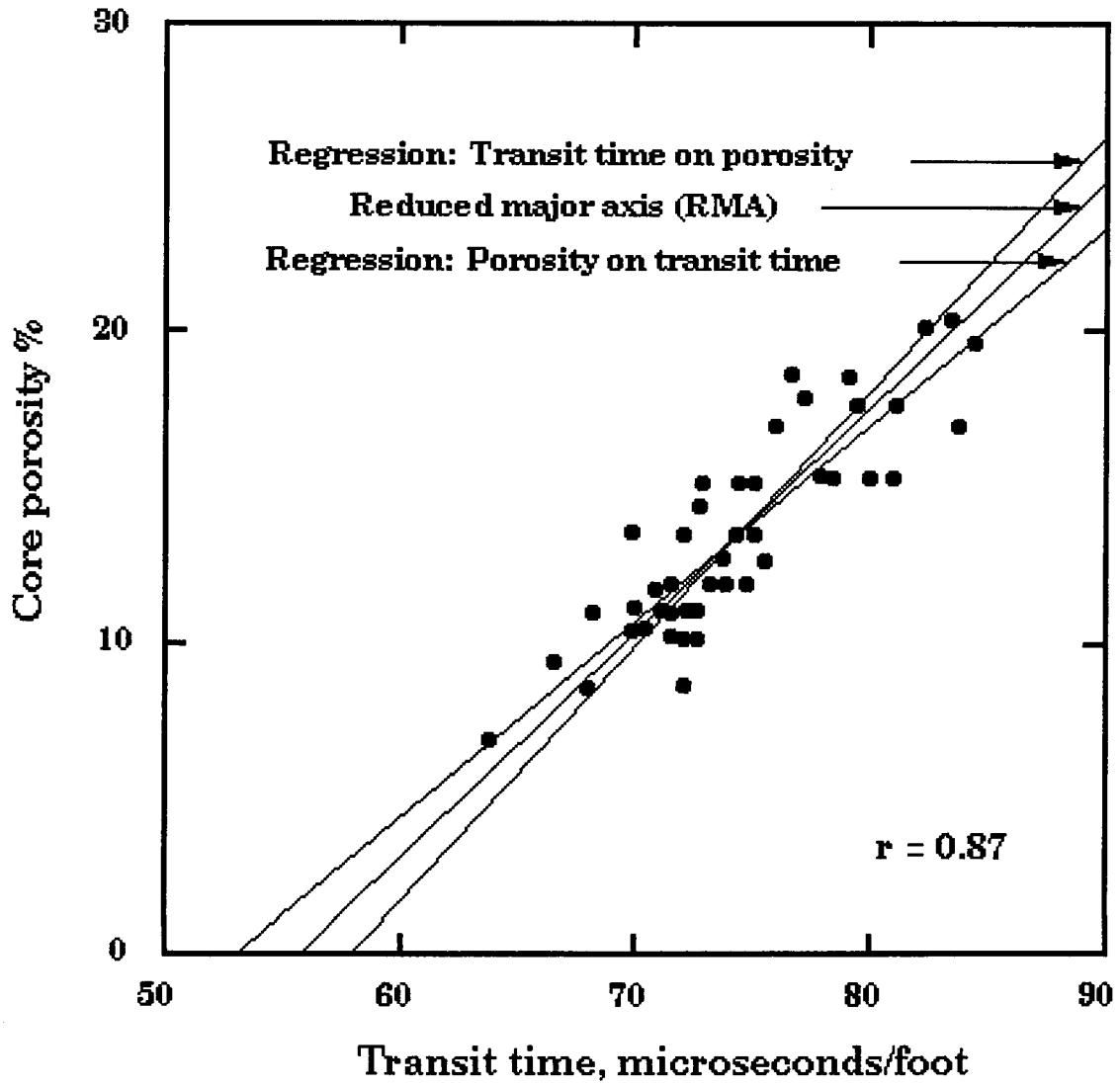
Φ	Δt	Φ	Δt
6.8	63.8	13.4	75.1
9.3	66.6	13.4	74.4
8.5	68.1	13.4	72.1
10.9	68.3	13.5	69.9
10.3	69.9	14.3	72.8
10.4	70.5	15.1	72.9
10.2	71.6	15.1	74.5
10.1	72.2	15.1	75.1
10.1	72.7	15.3	77.9
8.6	72.2	15.2	78.5
11.0	72.7	15.2	80.0
11.0	72.3	15.2	81.0
10.9	71.6	16.9	83.8
11.0	71.1	17.6	81.1
11.1	70.0	17.6	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 SADLEROCHIT SANDSTONE

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.

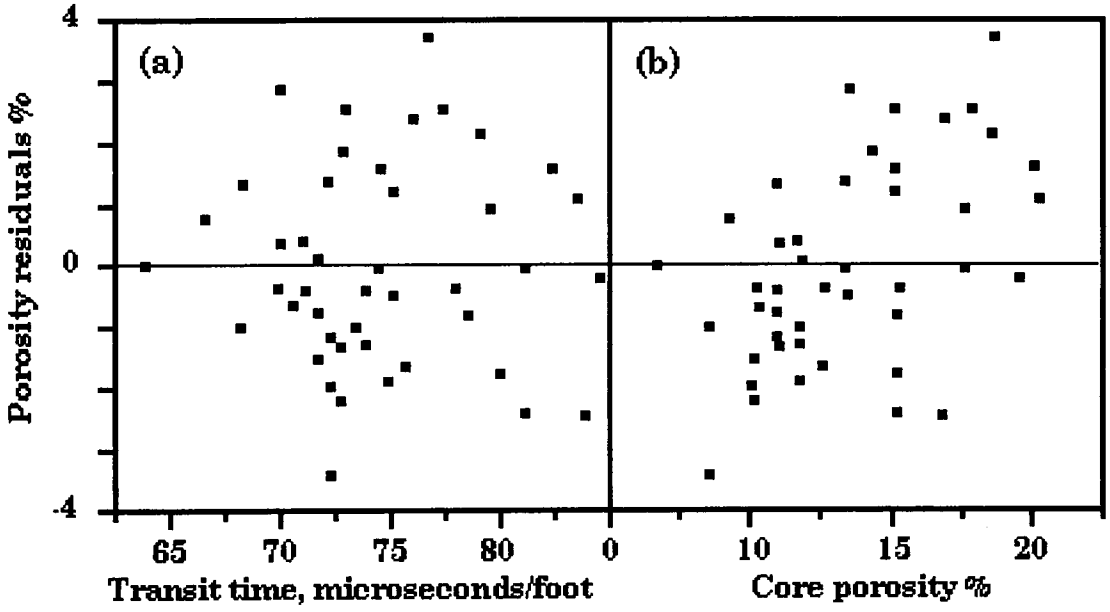


The regression equation is:

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

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 the plots below 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.



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.

The alternative regression of transit time on porosity results in the steepest line fit 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.

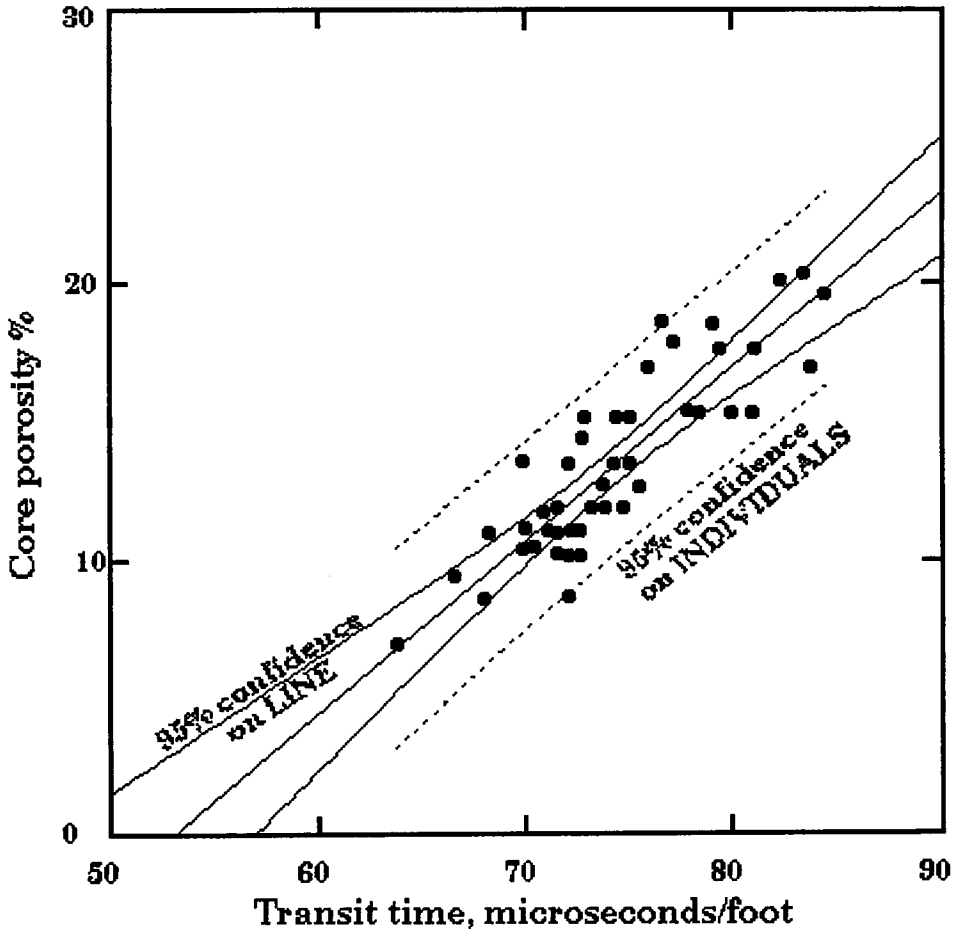
Finally, the prediction line that is often used as a compromise between the two regression extremes is the reduced major axis (RMA). The line passes through the bivariate mean and its slope is determined by the ratio of the standard deviations of the core porosity and transit time. The equation of the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

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 (the dashed lines on the plot below). If a 95% level is chosen, then we would expect 95% of the data points to plot within these bounds. The Sadlerochit 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. (A simple method to deal with outliers will be discussed in a later section.) 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 (the solid curves on the plot below).



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:

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

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:

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

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.

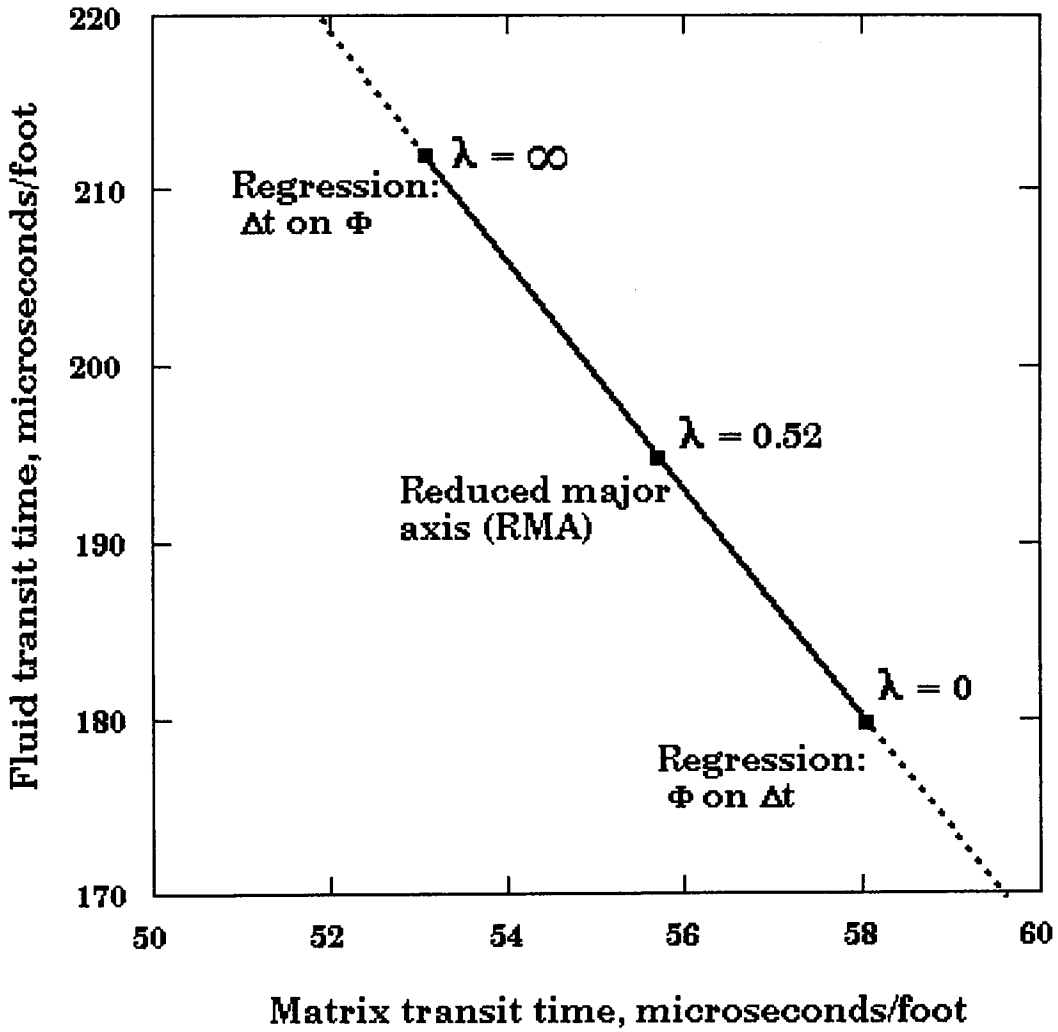
In summary, the choice of best-fit line is first determined by the purpose of the procedure. If the intent is only to make predictions of one variable on the basis of measurements of another, then regression is the preferred choice. The variable to be predicted is the dependent variable, the predictor 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 Sadlerochit 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

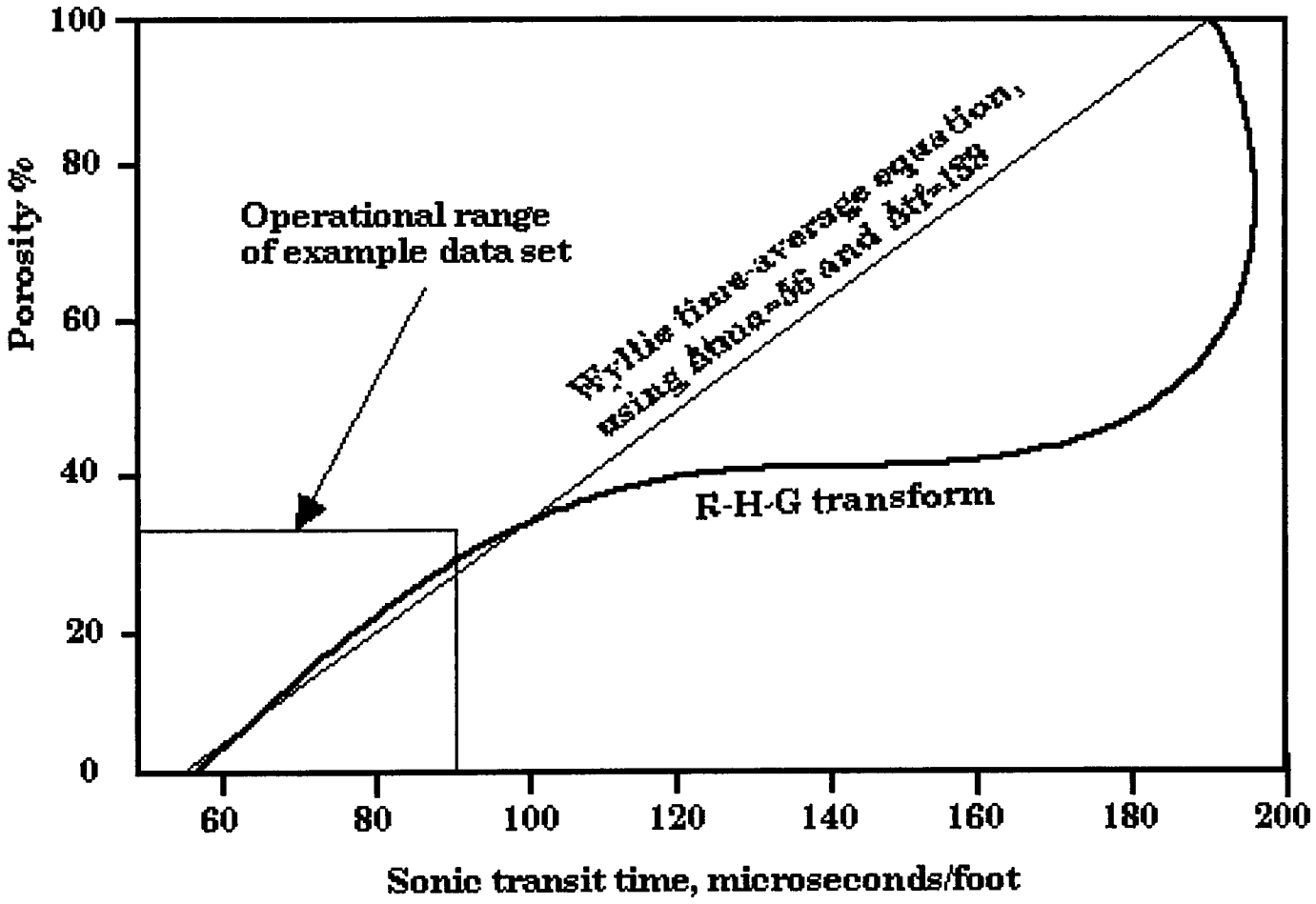
search of an optimum line-fit. The criteria to be met are that the joint estimates of the functional parameters and the error variance ratio, λ should be judged the most reasonable combination. The range of possibilities are bound by the regression line at each extreme, where the total error is attributed 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 below. 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.



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. 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.

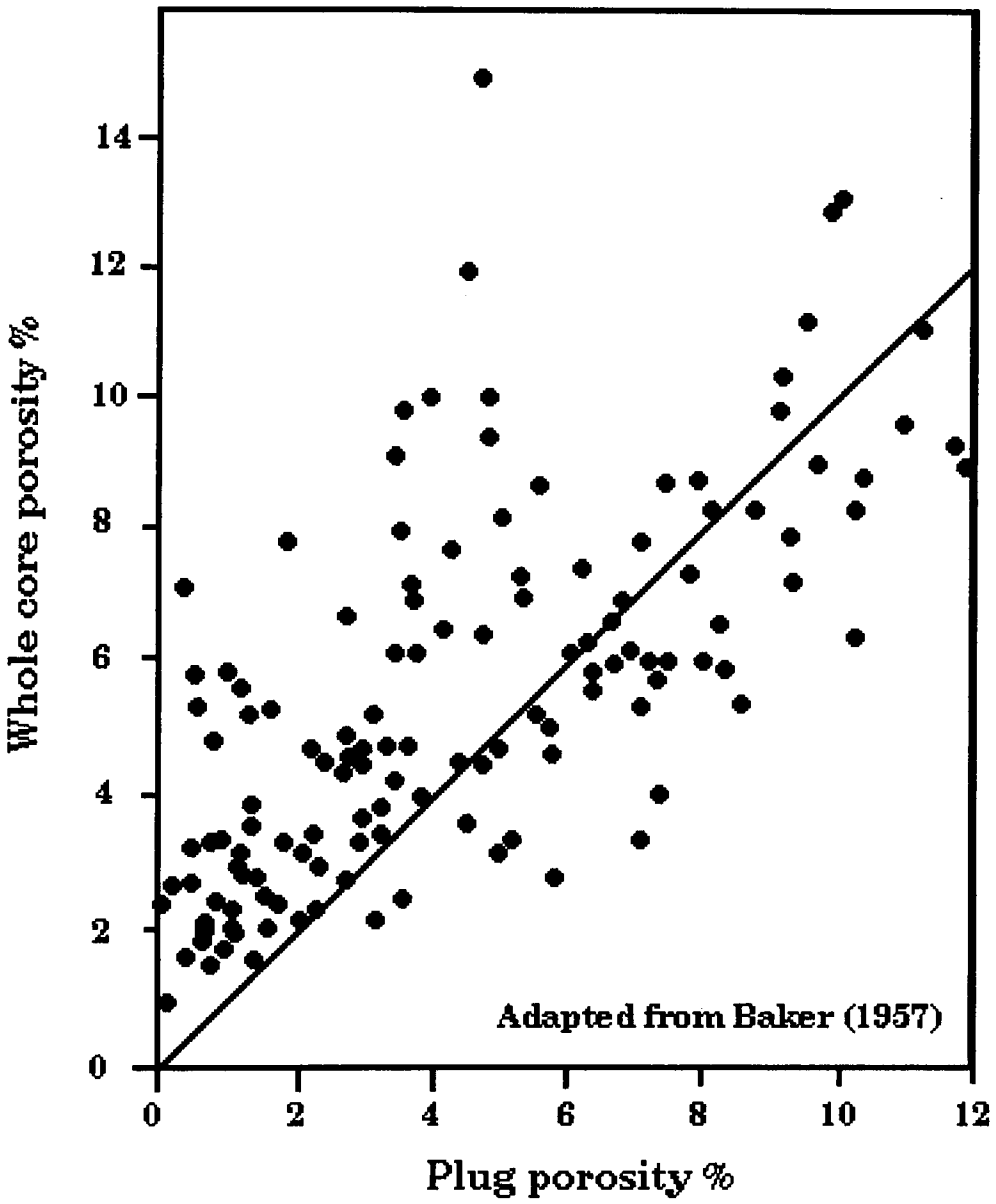
THE RAYMER-HUNT-GARDNER (R-H-G) SONIC TRANSIT TIME-TO-POROSITY TRANSFORM CONTRASTED WITH THE WYLLIE TIME-AVERAGE EQUATION



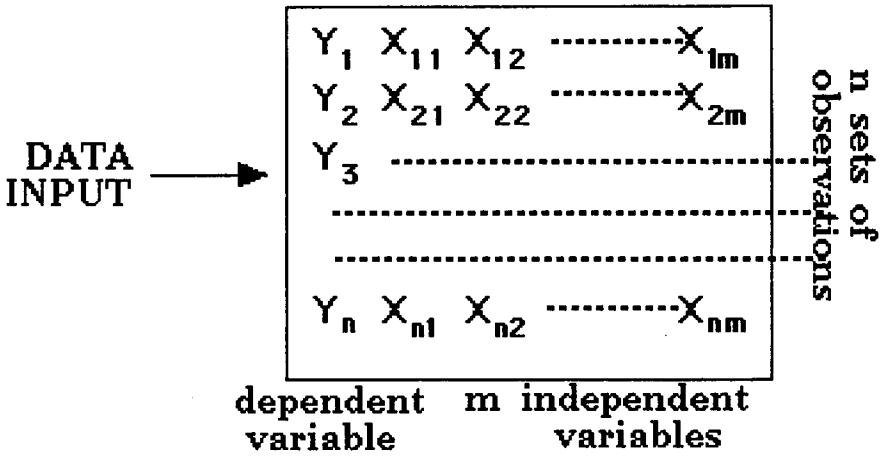
COMPATIBILITY OF VERTICAL RESOLUTION

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.


Whole core porosity %
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
Plug porosity %
Adapted from Baker (1957)

A dependent (or predicted) variable Y, is regressed on m independent (predictor) variables $X_1, X_2, \ldots X_m$
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 :

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

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

Now, the solution is found by minimizing the sum of squares deviations between Y and Y_i, given by :

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

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

These m equations rearranged in matrix form are :

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

$$SA = P$$

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

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

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

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 :

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

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 :

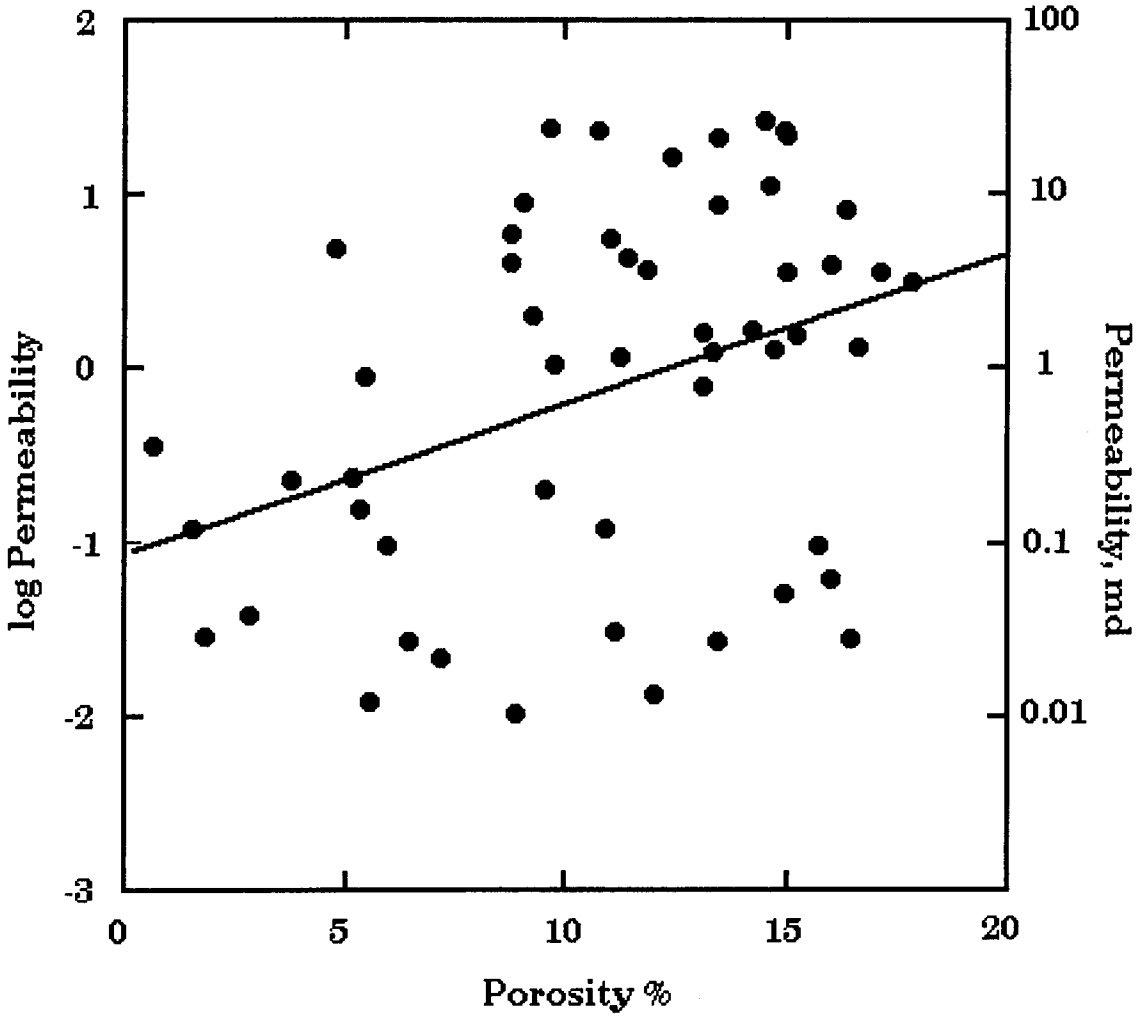
$$K = \frac{A\Phi^3}{(1-\Phi)^2 S^2}$$

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 below. 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.

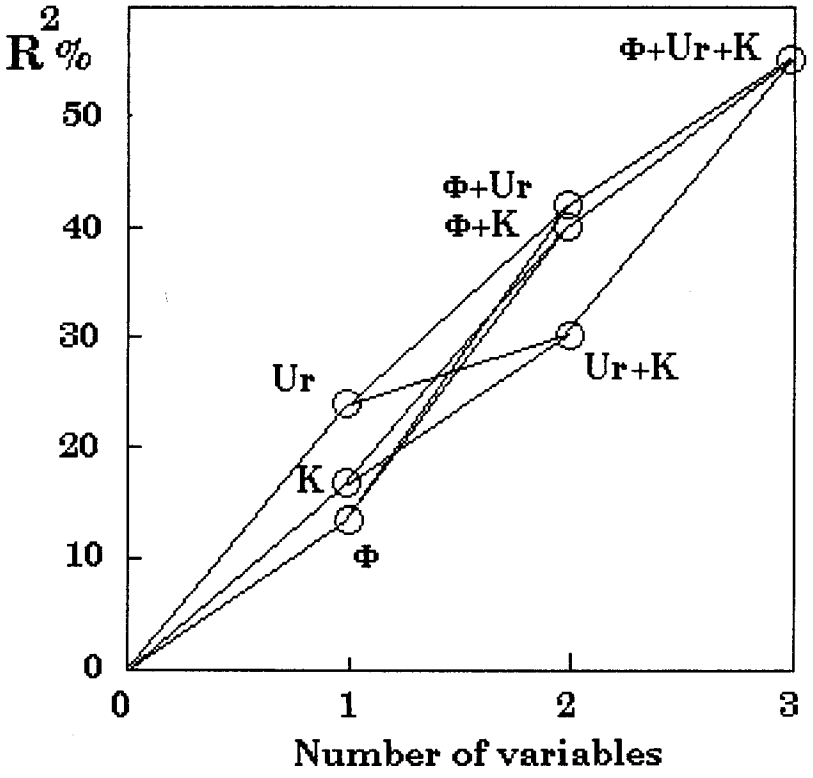


Multiple regression is an extension of simple linear regression analysis that incorporates additional independent variables in the predictive equation. By this means, permeability predictions may be improved through the inclusion of log measurements which are indirectly related with pore geometry, principally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

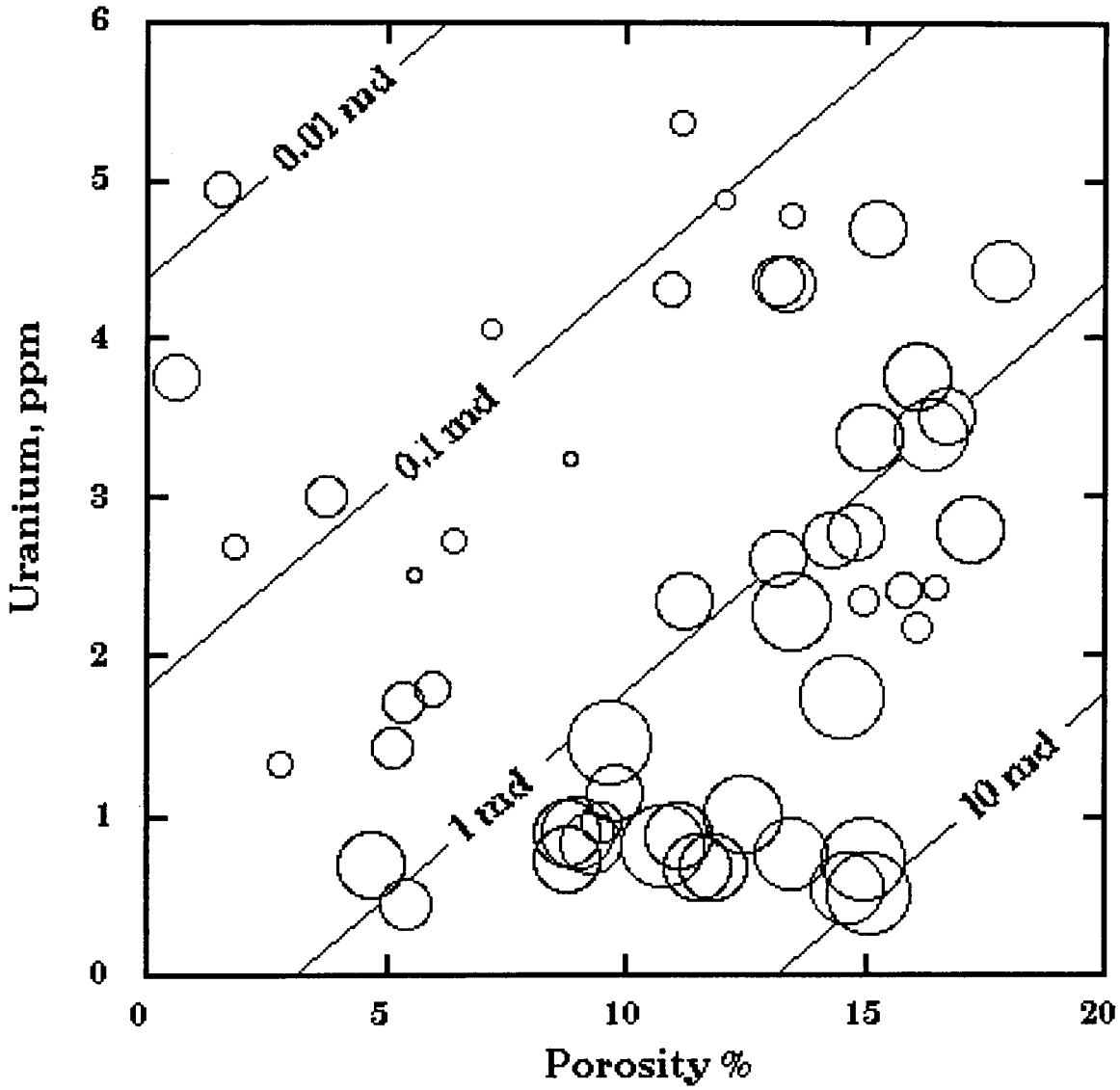
$$\log K = A + B * \Phi + C * L1 + D * L2 + \ldots$$

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.

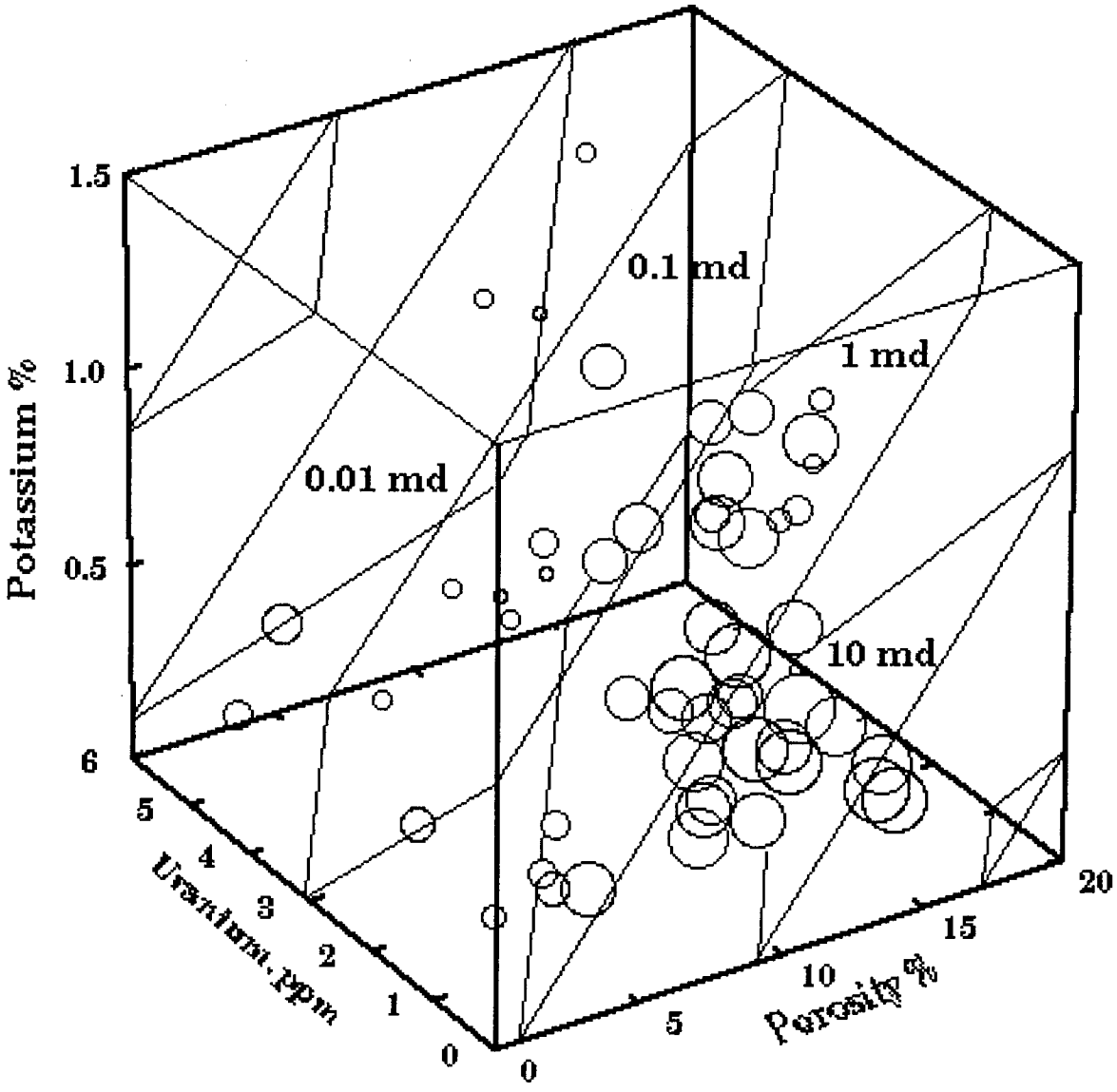
In the Chase Group example, the coefficients of determination for the alternative regressions of core permeability on all possible combinations of log measurements of porosity, uranium and potassium are shown below. There is a systematic improvement in prediction power with the inclusion of additional 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.



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.

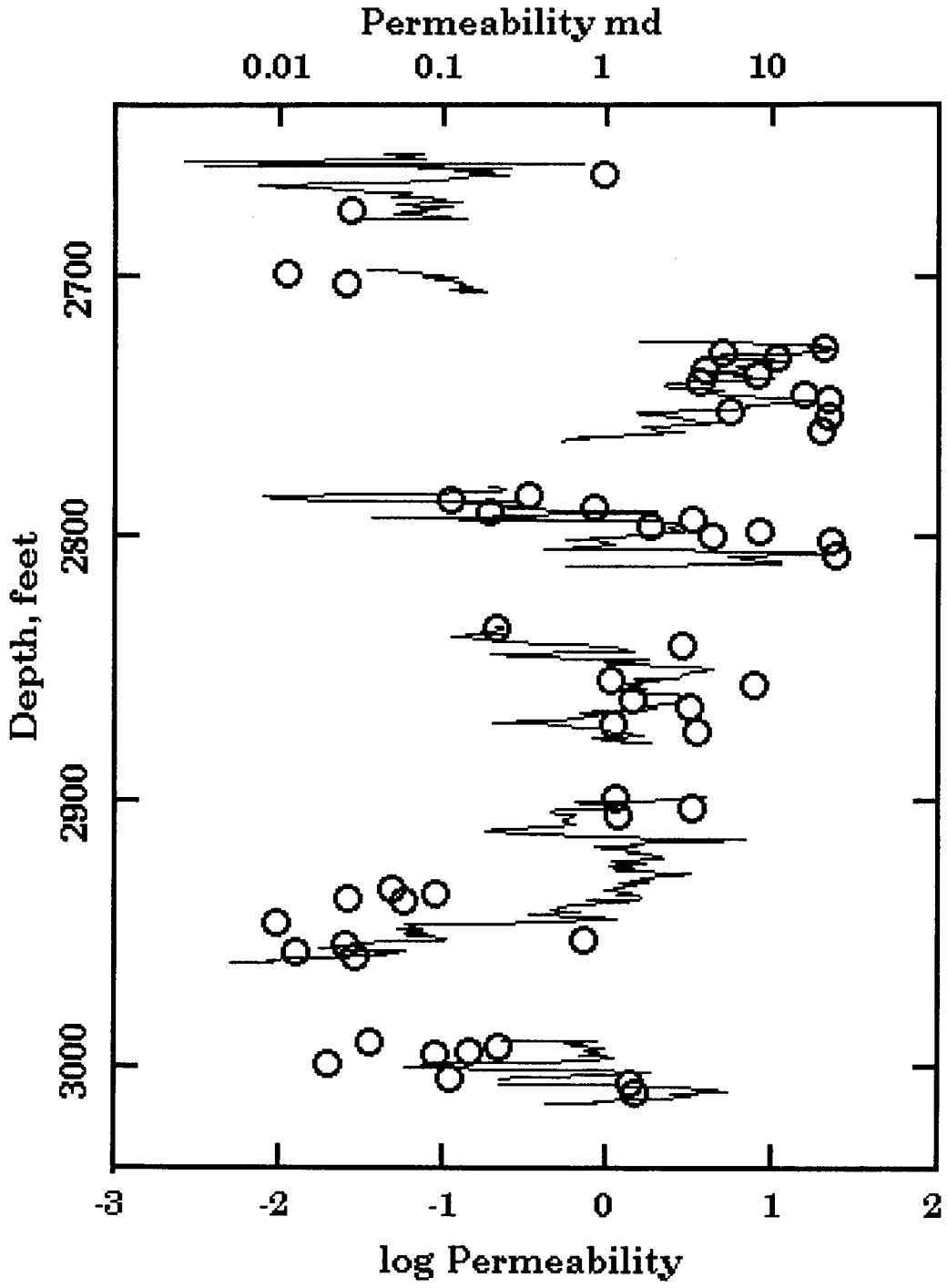


The regression coefficients associated with the independent variables show a consistent pattern of increasing permeability with increasing porosity, but decreasing permeability with greater concentrations of uranium and potassium. Both of these elements are statistically significant contributors to the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



Predicted permeability log of the Chase Group section from multiple regression on porosity, uranium, and potassium. Measured core permeabilities shown by circles.

POLYNOMIAL REGRESSION: AN EXAMPLE

Wireline logs record formation properties as a function of depth and can be fitted by curves generated by polynomial regression to compute major trends in property variation. A first-order polynomial is a straight line with the equation

$$\hat{S} = a_0 + a_1 d$$

where $\hat{S}$ is the regression estimate of the log at depth d, predicted by a line with intercept a_0 and slope a_1.

A quadratic (or second-order) polynomial function describes a curve with a single maximum or minimum and its application is given by the equation

$$\hat{S} = a_0 + a_1 d + a_2 d^2$$

The matrix algebra solution of the unknown coefficients is

$$\begin{bmatrix} n & \sum d & \sum d^2 \\ \sum d & \sum d^2 & \sum d^3 \\ \sum d^2 & \sum d^3 & \sum d^4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum S \\ \sum Sd \\ \sum Sd^2 \end{bmatrix}$$

The addition of extra polynomial terms at higher orders specify curves of increasing complexity. However, any polynomial order is fitted to raw data by the same generalized matrix algorithm. For an mth-order polynomial, the relationship is

$$\begin{bmatrix} n & \sum d & \sum d^2 & \sum d^m \\ \sum d & \sum d^2 & & \\ & & & \\ . & & . & \\ . & & . & \\ . & & . & \\ \sum d^m & & \sum d^{2m} & \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ . \\ . \\ . \\ a_m \end{bmatrix} = \begin{bmatrix} \sum s \\ \sum Sd \\ \sum Sd^2 \\ . \\ . \\ . \\ \sum Sd^m \end{bmatrix}$$

In a plot of regression fit against order, a polynomial can often be isolated that appears to satisfy a systematic major trend, while the residual variation is compounded from markedly finer-scaled fluctuations.

Polynomial functions have additional simple properties that provide useful parameters for mapping. If a polynomial equation of the form

$$\hat{s} = a_0 + a_1 d + a_2 d^2 + \cdots + a_m d^m$$

is differentiated, the result is

$$\frac{d\hat{s}}{dd} = a_1 + 2a_2 d \ + \cdots + m a_m d^{m-1}$$

which gives the slope of the polynomial at any depth. The location of the peaks and troughs of the polynomial can be extracted when the slope equation is set to zero and solved for depth. For example, when a cubic function is fitted to a log, the slope equation is a quadratic and yields two depth values as solutions. The depths correspond to the trend estimate of the maximum and minimum log property within the section.

A further differentiation of the generalized slope equation gives

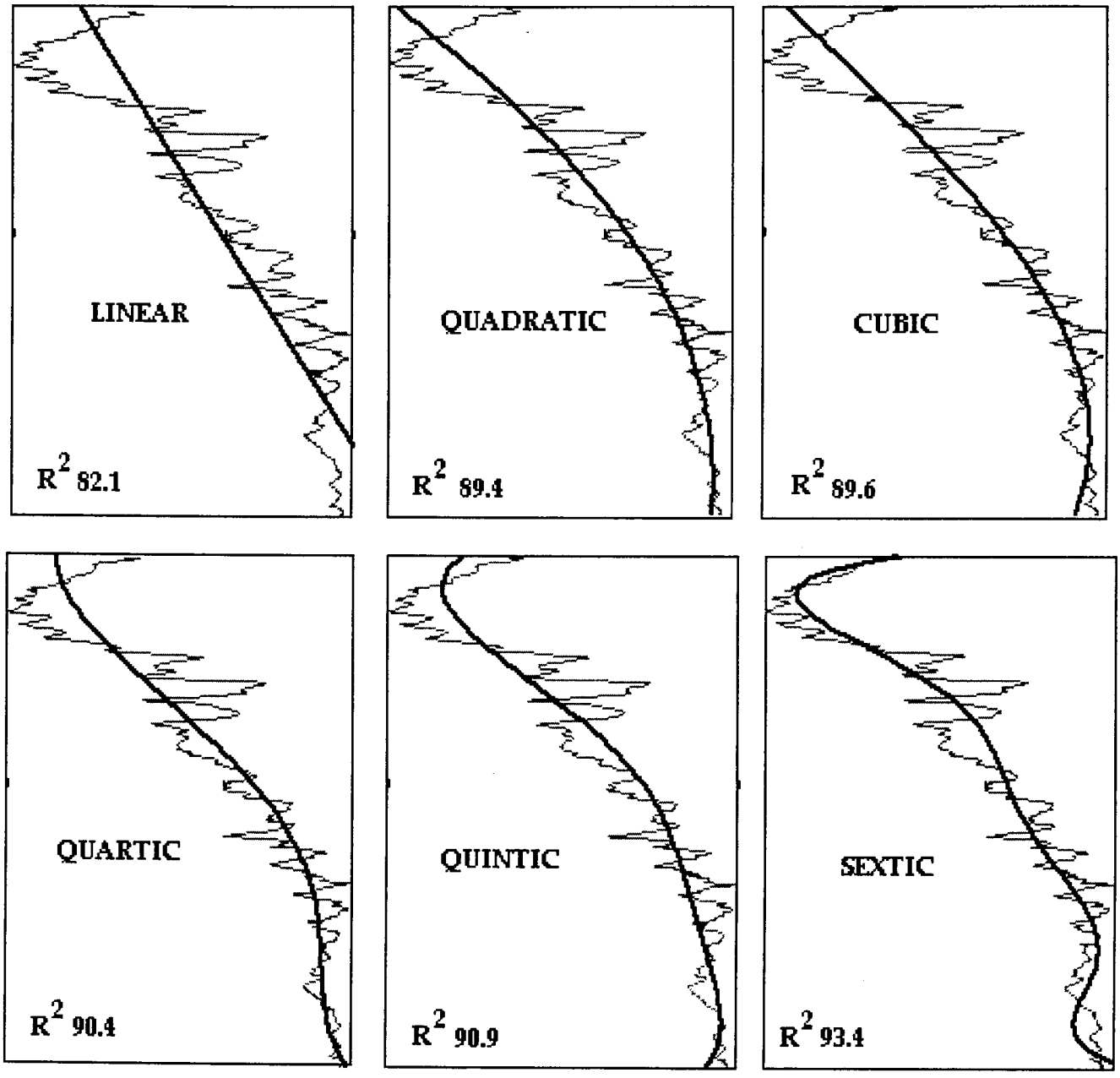
$$\frac{d^2\hat{s}}{dd^2} = 2a_2 + m(m-1)a_m d^{m-2}$$

which is an expression of the rate of change in slope. The rate of change is zero at curve inflection points, which mark the boundaries between peak and trough features.

In another application, the polynomial trends can be interpolated between well control because the polynomial functions are defined by moments. It follows that if we can interpolate moments between well control with some degree of confidence, then we can estimate those moments at undrilled locations and generate estimated polynomial trends. When the procedure is linked with a three-dimensional cell construct with geographic coordinate and depth axes, it would be possible to estimate the trend value of any cell as determined by the interpolated moments and their associated polynomial.

In the illusration, a hierarchy of polynomial curves are fitted to the porosity variation in the Mississippi Chat section.

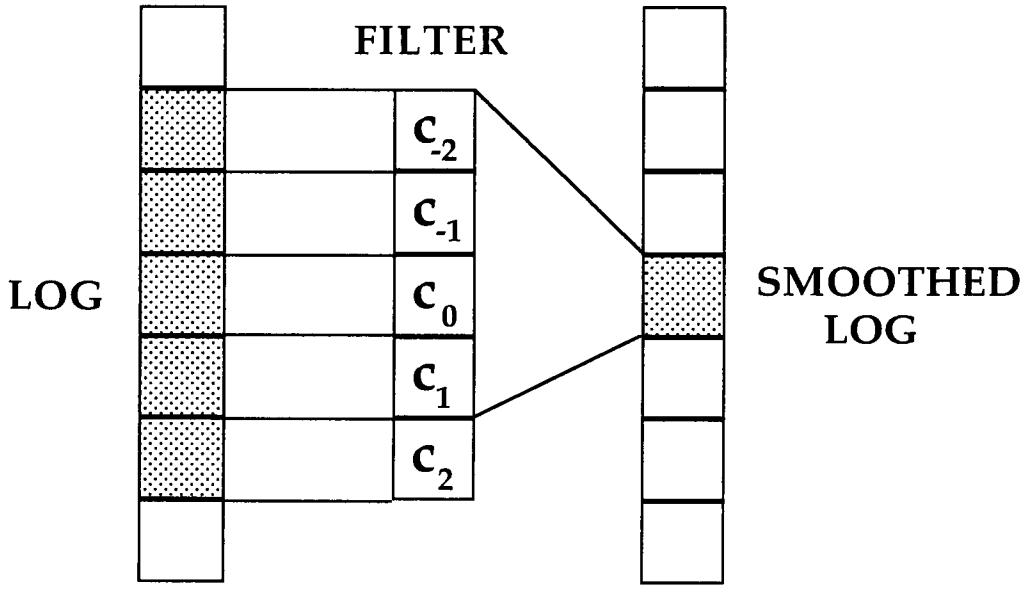
POLYNOMIAL REGRESSION OF POROSITY IN A MISSISSIPPI CHAT SECTION



POLYNOMIAL FILTERS

The polynomial regression described in the last section is a useful method to extract major trends of variation that reflect broad changes in rock or reservoir properties. Since the curve-fitting procedure utilizes all the data in the interval of interest, it is sometimes called a "global" procedure. By contrast, local methods are useful in the isolation of systematic intermediate and short-term trends through the analysis of interval subdivisions. Their operation results effectively in a smoothing of the raw data to a trace which is sensitive to localized trends.

The simplest method to smooth fluctuating data is to compute an equally-weighted moving average. The method is easy to program and requires only a specification of window length. If a log is averaged by a window of five feet, the smoothed result will screen out variations whose thickness is less than about half of this dimension, while retaining variation whose scale is greater. The operation is shown diagrammatically, where a column of digitized raw data is transformed by a moving average operator to a sequence of smoothed log values by successively sliding the operator past the log at incremental step positions. At each step, the operator elements are cross-multiplied with corresponding elements on the matched log segment and the results summed. The value is then divided by a normalizing factor, which in this case is the sum of the equal-valued operator elements, to obtain the smoothed estimate at the central position of the filter.



The arithmetic procedure of cross-multiplication and summation of products is known as "convolution". The moving average operator is an example of a "filter". The simple shape of this operator gives rise to its informal name as a "box car filter". In summary, the raw log variation is convolved by a filter to a smoothed trace.

The moving average is only one example of a convolution filter which can be used in the transformation of data sequences. Since its form is extremely crude, it does not extract as much information as more sophisticated filters. A better smoothing operation would be one which caused a best-fit curve to be drawn through a data segment scanned by the filter window. If the criterion of best-fit is set by the principle of least squares and the curve represented by a polynomial function then the goal is a localized polynomial regression. A polynomial that is commonly used in filtering is the cubic, because the equation will accommodate up to two turning points (a maximum and a minimum) while also describing simple trends or one turning point (a maximum or minimum).

Polynomial curve fitting involves the calculation of sums of powers and cross-products of two descriptive variables in a matrix format, with matrix operations of inversion and multiplication. This procedure would be extremely cumbersome to apply at each successive step in the migration of the window down the length of the log trace. However, an equivalent computation can be achieved through the design of a convolution filter which takes advantage of the fact that successive values of depth are separated by a fixed increment on a digitized record. The approach was first described by Savitzky and Golay (1964) and polynomial filters are often known as "Savitzky-Golay filters".

The principles of polynomial filter design are explained with reference to a cubic function which is specified by the equation:

$$s = a_0 + a_1 d + a_2 d^2 + a_3 d^3$$

where s is the cubic trend in log response and d is depth position. The filter window is made up of $(2m + 1)$ elements ranging in depth value from $-m$ to $+m$, with a value of zero at the midpoint. At the location of the window midpoint (where the depth value is zero), the value of the cubic function is given by a_0, while differentiation of the cubic equation gives a first derivative of a_1 and a second derivative of $2a_2$ The standard matrix algebra solution for the coefficients of a cubic regression model can be written as:

$$\begin{bmatrix} n & \sum d & \sum d^2 & \sum d^3 \\ \sum d & \sum d^2 & \sum d^3 & \sum d^4 \\ \sum d^2 & \sum d^3 & \sum d^4 & \sum d^5 \\ \sum d^3 & \sum d^4 & \sum d^5 & \sum d^6 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \sum s \\ \sum sd \\ \sum sd^2 \\ \sum sd^3 \end{bmatrix}$$

where n is the number of filter elements and the summations are over all $(n = 2m + 1)$ elements. The system collapses to a simpler equation set, because all terms that involve the summation of depth values with odd exponents are zero.

The cubic function value at the midpoint is therefore given by:

$$\frac{\sum d^4 \sum s - \sum d^2 \sum sd^2}{n\sum d^4 - \sum d^2 \sum d^2}$$

For a nine-element filter, the loadings are solved as:

[-21 14 39 54 59 54 39 14 -21]

with a normalizing factor (the divisor that is applied, following cross-multiplication and summation) of 231.

For a 25-element filter, the loadings are:

[-253 -138 -33 62 147 222 287 343 387 422 447 462 467 462 447 422 387 343 287 222 147 62 -33 -138 -253]

with a normalizing factor of 5175.

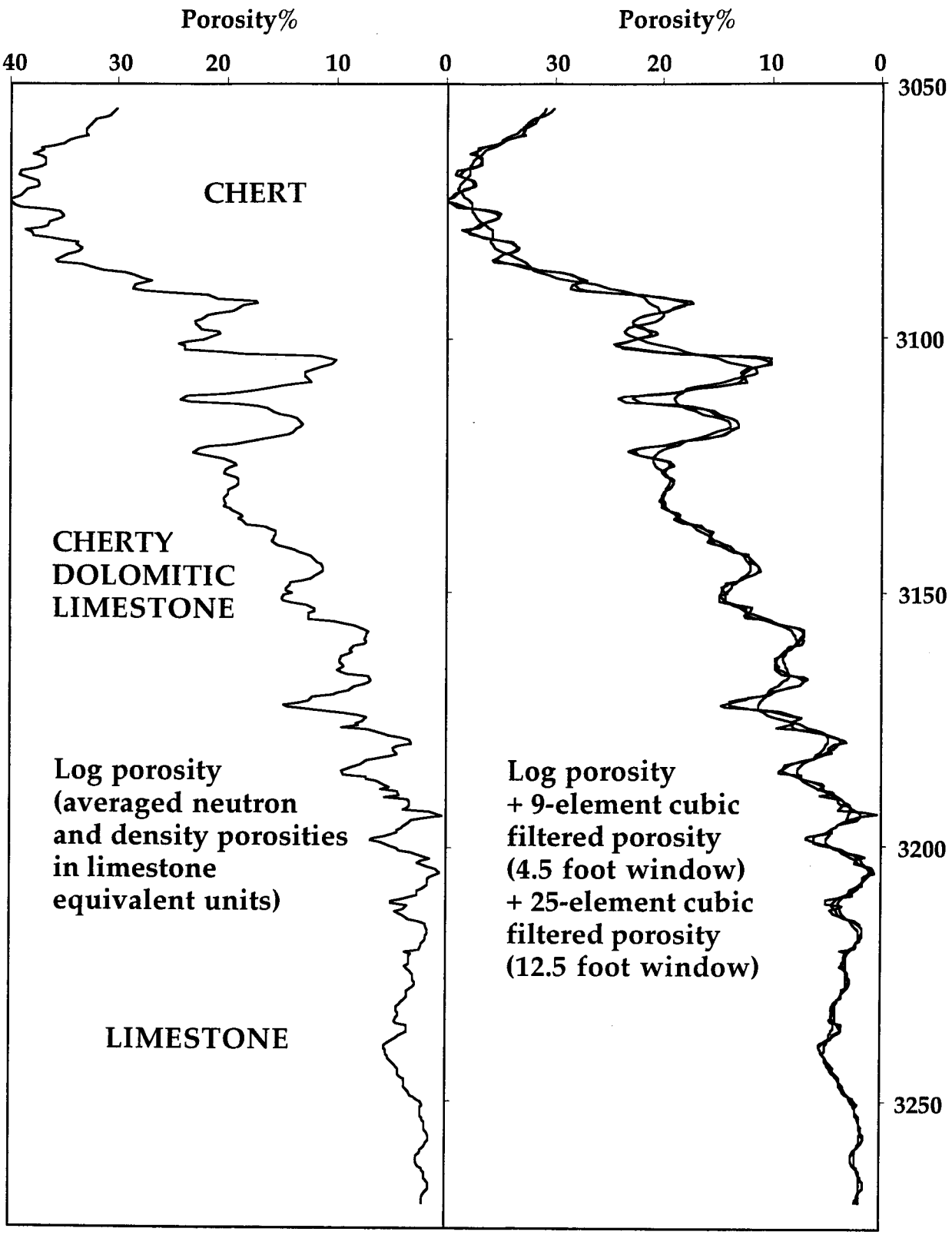
A cubic polynomial is often used as a useful function which would be adequate to fit peaks, troughs, or shoulders at the scale of the window span. A table of loadings for cubic polynomial Savitzy-Golay filters for windows of various sizes is listed, where NP is the number of elements, the values of a are the loadings referenced to their position in the filter, and h is the normalizing factor. (Because the filters are symmetrical, the loadings for the negatively subscripted elements take the same values as their positive equivalents.)

NP	h	a0	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12
5	35	17	12	-3	0	0	0	0	0	0	0	0	0	0
7	21	7	6	3	-2	0	0	0	0	0	0	0	0	0
9	231	59	54	39	14	-21	0	0	0	0	0	0	0	0
11	429	89	84	69	44	9	-36	0	0	0	0	0	0	0
13	143	25	24	21	16	9	0	-11	0	0	0	0	0	0
15	1105	167	162	147	122	87	42	-13	-78	0	0	0	0	0
17	323	43	42	39	34	27	18	7	-6	-21	0	0	0	0
19	2261	269	264	249	224	189	144	89	24	-51	-136	0	0	0
21	3059	329	324	309	284	249	204	149	84	9	-76	-171	0	0
23	805	79	78	75	70	63	54	43	30	15	-2	-21	-42	0
25	5175	467	462	447	422	387	343	287	222	147	62	-33	-138	-253

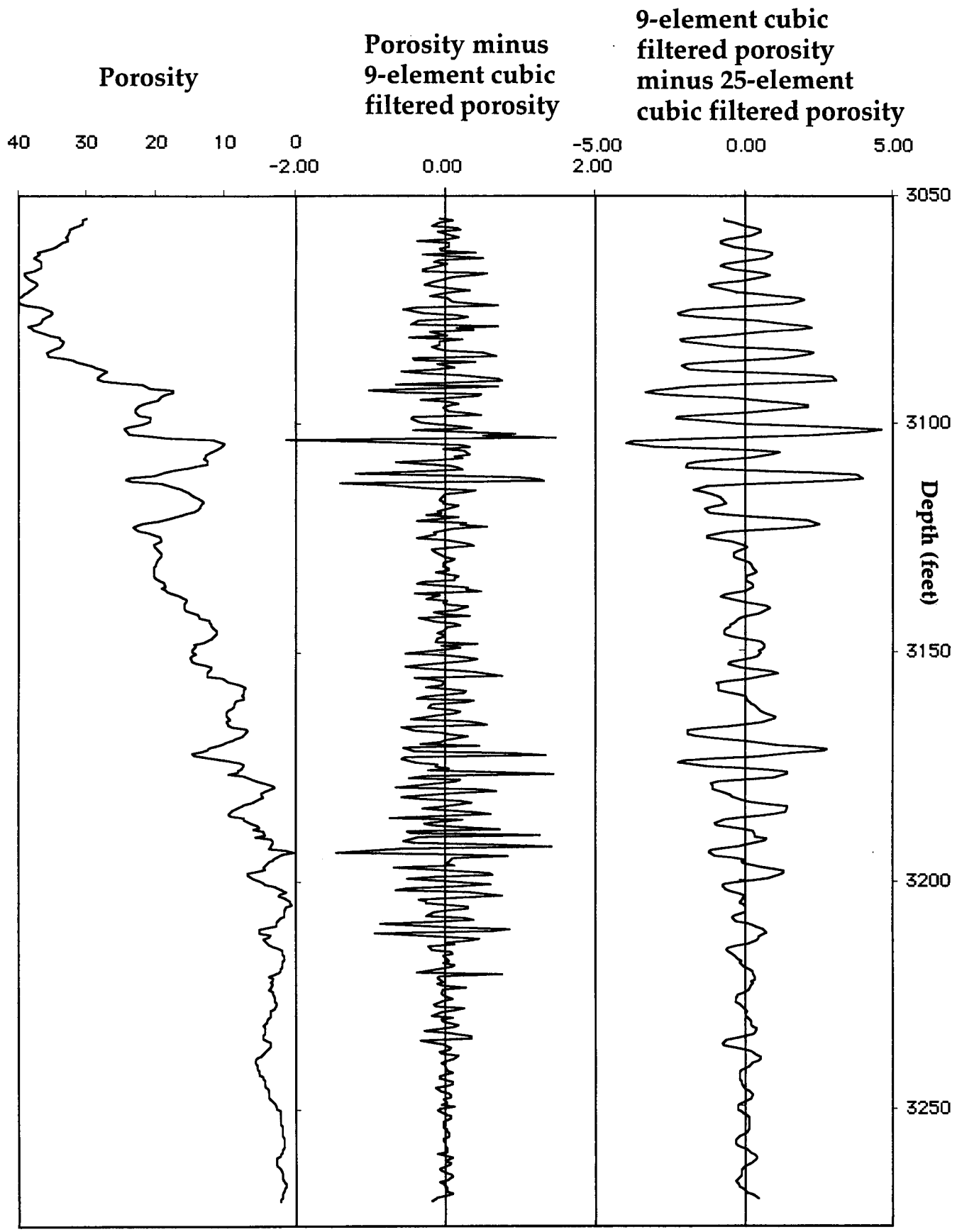
APPLICATION OF SAVITZKY-GOLAY POLYNOMIAL FILTERS TO A POROSITY LOG

The "Mississippi Chat" is the informal name given to an unusual lithology that occurs in the subsurface of southern Kansas and forms the reservoir unit for a number of important gas fields. The "Chat" is dominated by spiculitic chert whose porosity can range up to 50%, most of which consists of micropores from the dissolution of sponge spicules. The porosity log used in this example was taken from a Mississippian section in well in Sedgwick County, Kansas The section is capped by spiculitic chert underlain by dolomitic cherty limestone, and cherty limestone which becomes progressively less cherty with depth. The changes in lithology with depth are reflected in the changes in porosity. The estimate of porosity was made using neutron and density logs and the digital porosity log was created at the standard frequency of two readings per foot of depth.

A comparison of the cubic smoothing of the nine- and 25-element Savitzky-Golay filters with the original log does not show much difference at the large scale because of the small window lengths (4.5 and 12.5 feet) However, when the 9-element smoothed log is subtracted from the original log, the resulting trace shows porosity features at a scale of about two-feet or less. Subtraction of the 25-element smoothed log from the nine-element smoothed log gives a trace of porosity features whose thickness ranges approximately between two feet and six feet. The subtraction of the results of two filters is an example of band-width filtering where features within a certain range are displaced. The contrasts with the results from a single filter which are often characterized as "low-pass" or "high-pass" filters depending on whether they emphasize low-frequency or high frequency components in the original data.



Mississippian section porosity log together with smoothed logs generated by 9-element and 25-element Savitzky-Golay cubic polynomial filters



Mississippian section porosity with filtered porosity residuals

POLYNOMIAL DERIVATIVE FILTERS

At the location of the window midpoint (where the depth value is zero), the value of the cubic function is given by a_0 , while differentiation of the cubic equation gives a first derivative of a_1 and a second derivative of $2a_2$
The solution for the first derivative at the window midpoint is then:

$$\frac{\left(\sum d^4 \sum d^3 s - \sum d^6 \sum ds\right)}{\left(\sum d^4 \sum d^4 - \sum d^2 \sum d^6\right)}$$

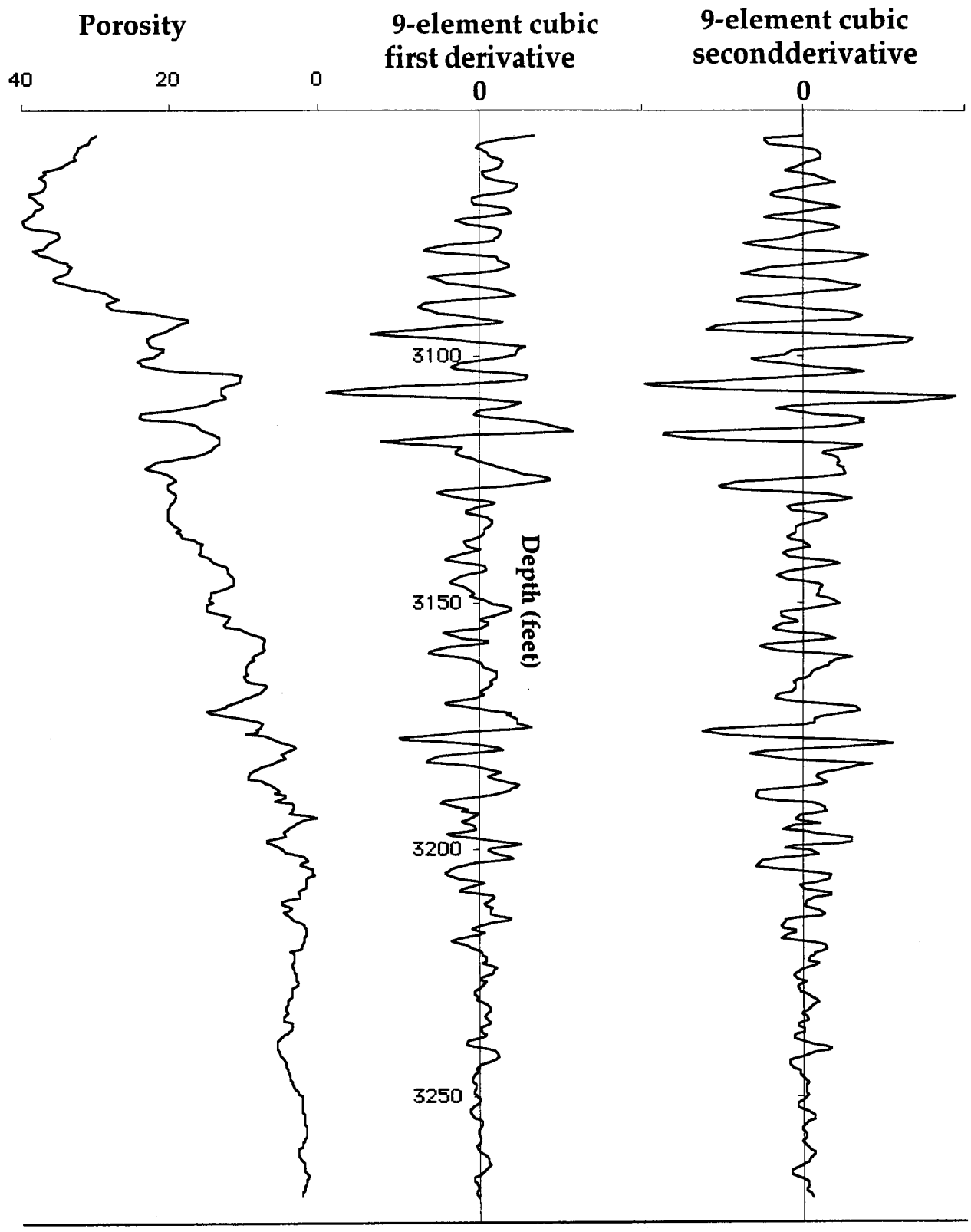
and the second derivative is:

$$\frac{2\left(n\sum d^2 s - \sum d^2 \sum s\right)}{\left(n\sum d^4 - \sum d^2 \sum d^2\right)}$$

For a cubic Savitzky-Golay filter with nine elements, the loadings for the first derivative filter are:

and for the second derivative are:

Application of these filters to the Mississippian porosity log are shown in the figure. The first derivative reflects slope or change in porosity values in the 4.5 foot window, where depth locations matched with zero values isolate maxima and minima discriminated by the localized cubic fitting. The second derivative measures rate of change of porosity and depths of zero values match inflection points as boundaries between porosity features filtered by this window.



Mississippian section porosity log with 9-element cubic polynomial filter first and second derivatives

GLOBAL ZONATION METHODS

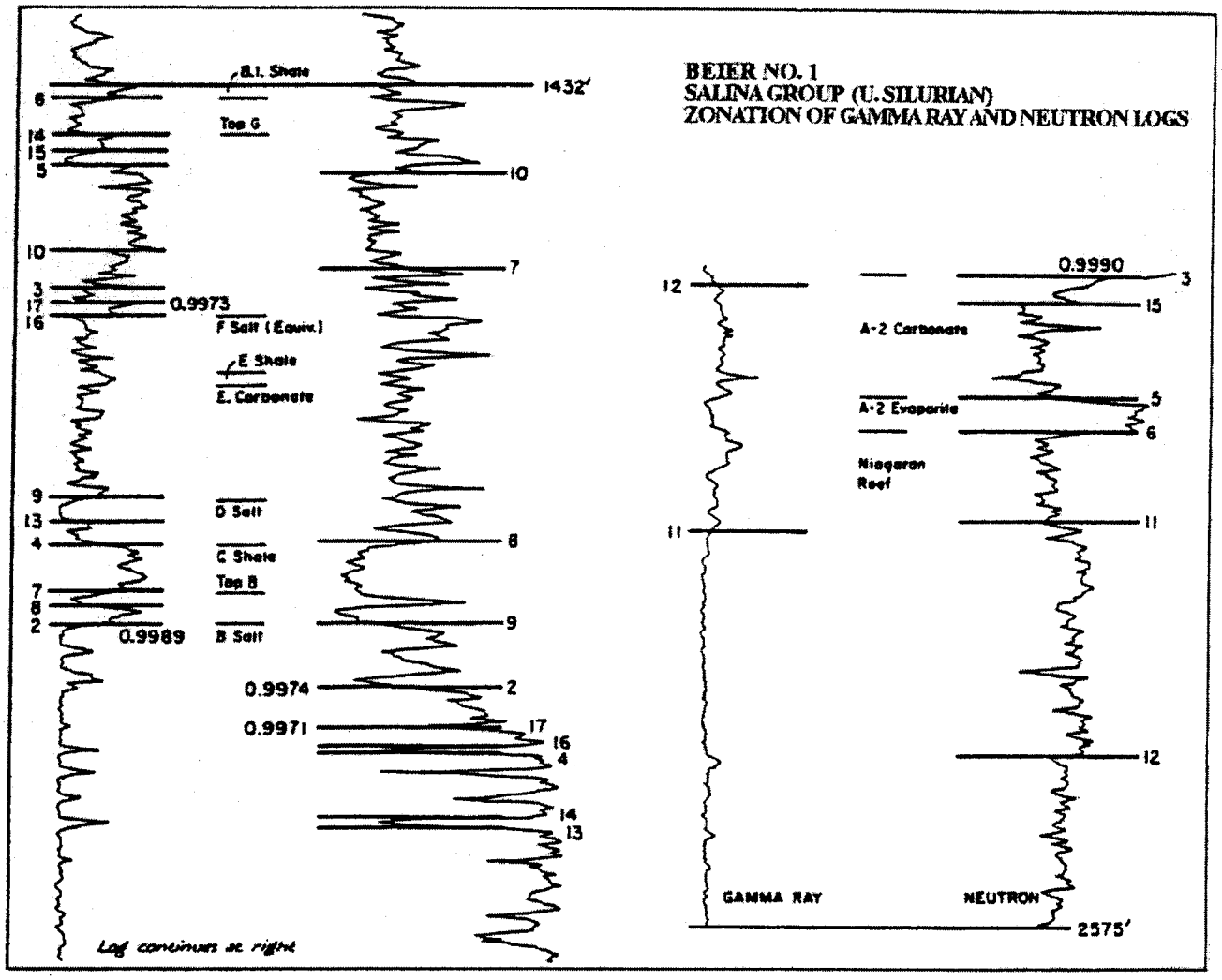
When subsurface geologists have located the tops of formation units on logs, they often subdivide the formations into lithostratigraphic units. Between the boundaries, each unit should be relatively homogeneous internally and show a marked difference with adjacent units. In their examination of a reservoir section, petroleum engineers pursue a similar goal in their subdivision of the section into distinctive flow units whose hydraulic properties are relatively homogeneous within each unit, but differentiated from vertically adjacent units.

The process of subdividing a formation or reservoir is usually called either "segmentation" or "zonation". Some methods apply a filter to hunt for boundaries between units but the results depend on the size of the filter window and are 'local' as contrasted with 'global' methods that consider the data from the entire section during the analysis.

GLOBAL VARIANCE PARTITIONING

The most common methods for log segmentation define subdivisions using either a variance or a mean-value criterion. Gill (1970) devised a computer methods that minimized variance within segments, while maximizing variance between them. Their procedures are adaptations of a single-factor, fixed-effect analysis-of-variance model. On the first pass, a boundary is moved progressively down the succession in an iterative series of trials that contrasts the variance within and between the two resulting segments. The boundary with the best differentiation marks the interface between the first two segments. The operation is then applied to each of the segments, and the process repeated in progressively finer segmentation. Either the number of segments must be known beforehand or some type of variance criterion must be applied to define a stopping point for the procedure.

An example of this method is from Gill (1970) shows the segmentation of gamma-ray and neutron logs in an Upper Silurian Salina Group section of carbonate, evaporite, and shale units from Michigan. Notice the numbers marked on the boundaries that register the order of their appearance in the segmentation procedure. This information can be useful in the weighting of the relative importance of the boundaries as markers for stratigraphic correlation. The segmentation follows a different hierarchy for the two logs because of their sensitivity to different petrophysical properties. Gill (1970) noted that the segments defined from the gamma-ray log were compatible with lithostratigraphic units picked manually. However, the association was not so close for segments partitioned from the neutron log, which responds primarily to porosity variability. Either of these segmentation results could be the "correct" result, depending on whether the end product should be a stratigraphic section keyed to lithologies, or a reservoir subdivided into distinctive porosity units.

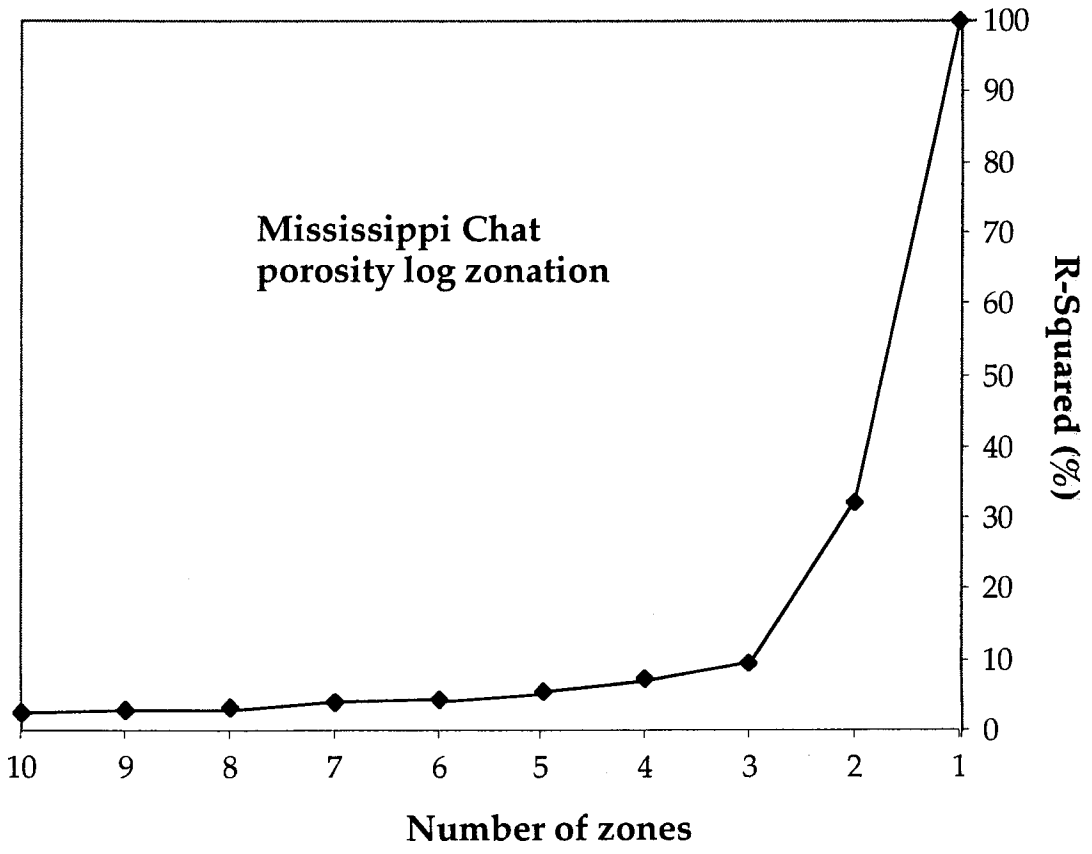


Segmentation of gamma-ray and neutron logs in an Upper Silurian Salina Group succession of carbonates, evaporites, and shales in a Michigan well. The numbers signify the order in which the segments were picked. From Gill (1970).

GLOBAL AGGLOMERATION: DEPTH-CONSTRAINED CLUSTERING

More recently, Gill et al. (1993) applied a adjacency-constrained cluster analysis to the segmentation problem. Unlike the divisive strategy of traditional variance techniques, the method is agglomerative in building coarser segments from finer divisions. A cluster dendrogram display is helpful in the search for possible natural levels in the hierarchy that may reflect meaningful geological boundaries. Finally, the method can be used with any number of logs, because it operates on similarity measures between adjacent zones or segments. Comparison of different subdivision results can be made in terms of analysis of variance, similar to the partitioning method in an evaluation of how much of the total variance is absorbed by the subdivisions.

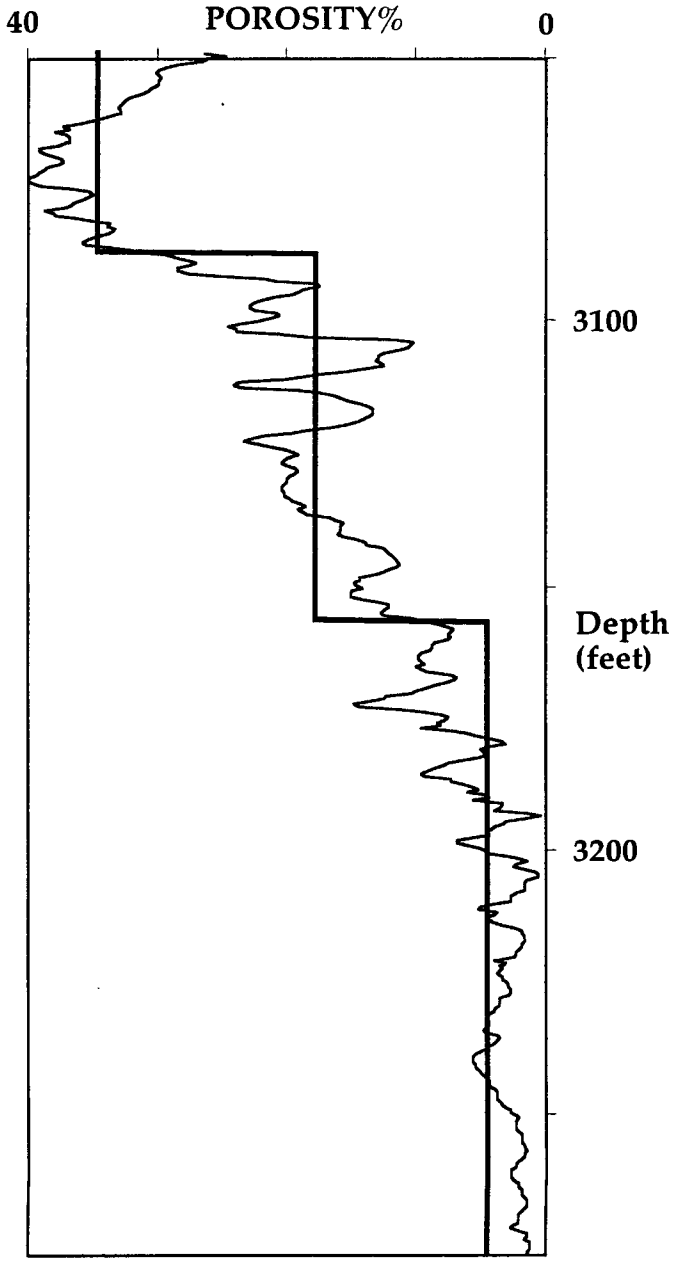
The porosity log from the Mississippi Chat section used to demonstrate the operation of the Savitzsky-Golay polynomial filters will be used as an example. The graph plots the R-squared value from the analysis of variance associated with each level of zonation which is the ratio of the sums-of-squares within the zones divided by the total sums of squares or: $\|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2$ At one extreme, a single zone for the entire section contains the total variability and R2 is 100%. At the other extreme, if each depth increment corresponds to a zone, then all the variability is between the zones and none within. The graph indicates that a subdivision into three zones is a fundamental descriptor. The associated ANOVA table is shown on the next page.



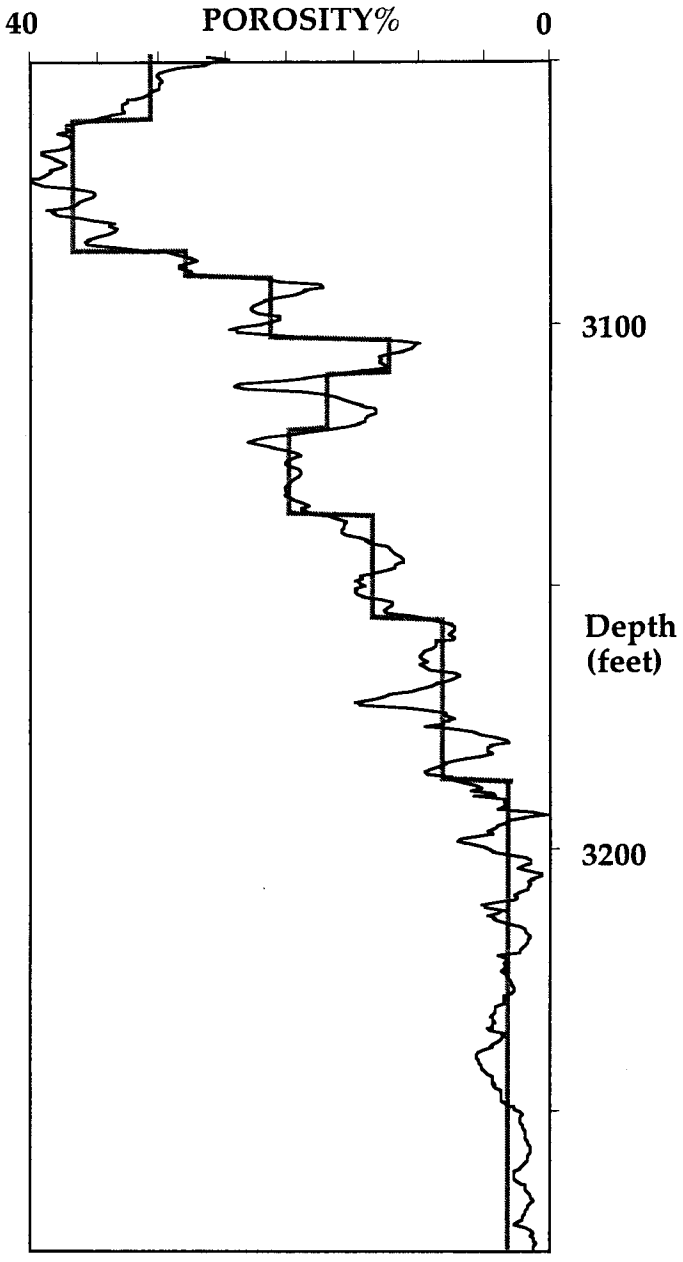
Group	Count	min. dep.	max. dep.	mean	s.d.	SSG	SSW	
1	76	3049	3086.5	34.63	3.65	998.2	5899.0	
2	139	3087	3156	17.74	4.68	3022.9		
3	239	3156.5	3275.5	4.43	2.81	1877.9		
							SSB 56072.9	
ALL	454	3049	3275.5	13.56	11.70		SST 61971.9	R2% 9.51881227

ANOVA table for agglomeration of Mississippi Chat section log porosity into three zones

The corresponding segmented porosity log for three zones is:



For comparison, the segmented porosity log for ten zones is:



The choice of the appropriate zonation level should be made on both statistical and functional criteria. Remember that the statistics will be sensitive to average properties as contrasted with localized features that may have major functional importance to the reservoir engineer or geologist.

ZONATION WITH MULTIPLE VARIABLES

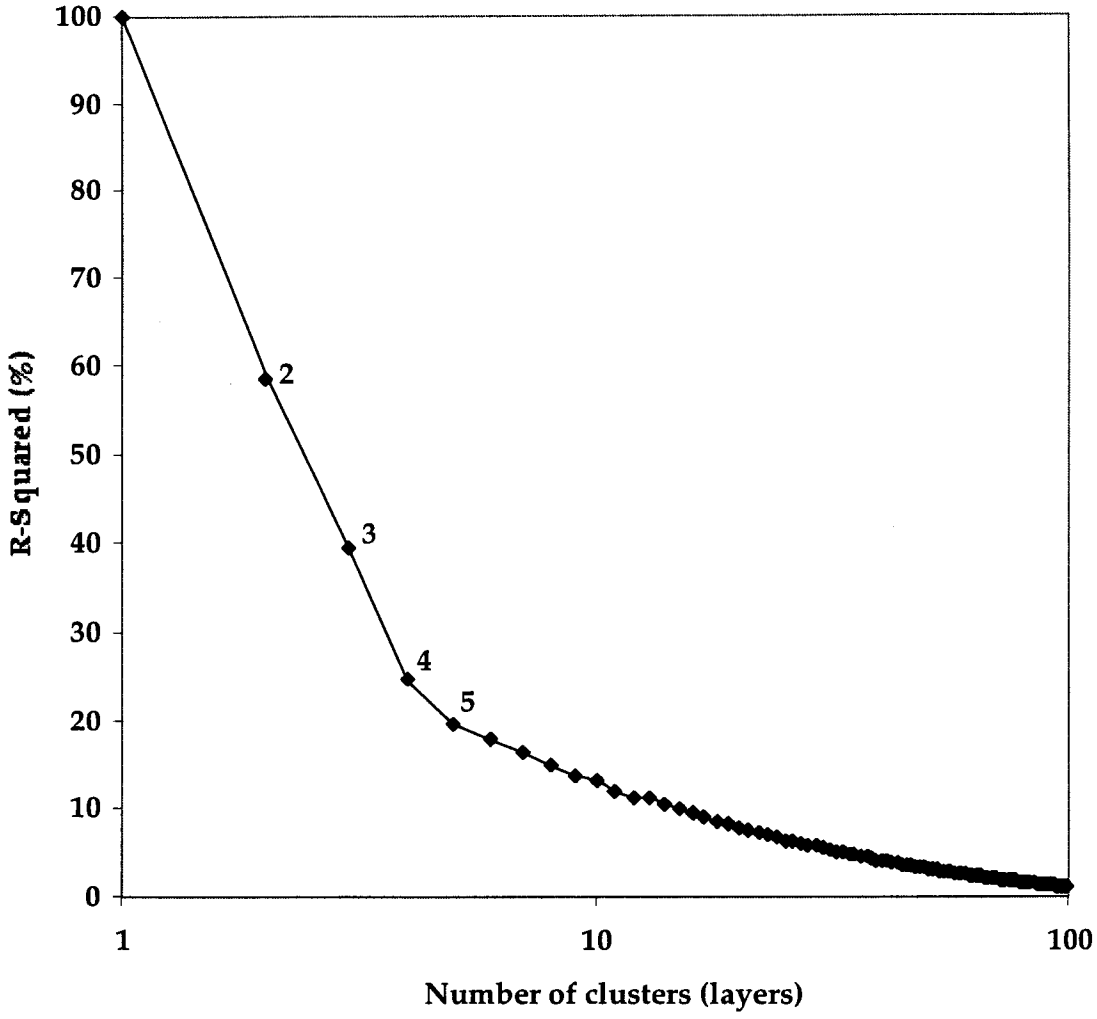
Depth-constrained clustering can be extended beyond a single variable such as porosity. The input logs could be the raw measurements, compositional analysis logs, or other measurements, according to the needs and purpose of the study. In this example, we extend the analysis of the Mississippi Chat section to include its composition of proportions of chert, dolomite, calcite and porosity estimated from neutron, density, and photoelectric factor logs.

Each of the logs employed is first standardized to zero mean and unit standard deviation before clustering, in order to ensure that they all have approximately equal weight in the analysis. The clustering employs Ward's method, which, at each step of the process, joins the two groups (subintervals) that are most alike in a least-squares sense. That is, it joins the two groups whose merger produces the least possible increase in the total within-groups sum-of-squares. The sum-of-squares for a single group, k, is given by

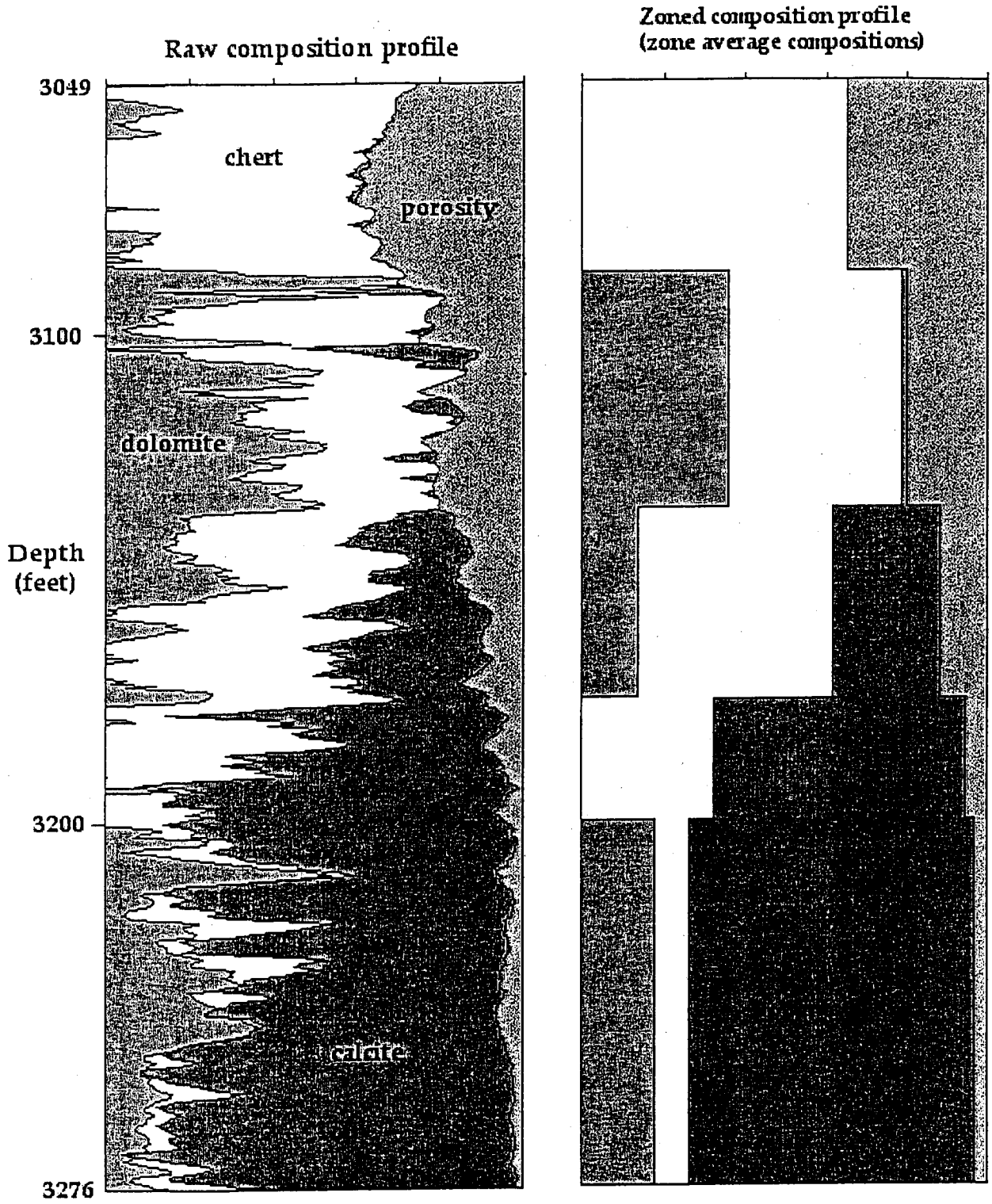
$$W_k = \sum_{i=1}^{n_k} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2$$

where $\|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2$ is the squared distance between the vector of (standardized) log values for data point i, $\mathbf{x}_i$, and the vector mean for group k, $\bar{\mathbf{x}}_k$. The within-groups sum-of-squares, W, is simply the sum of the W_k values over all groups. At each step of the clustering process, the number of groups is reduced by one and the within-groups sum-of-squares increases. Depth-constrained cluster analysis only allows vertically adjacent groups (subintervals) to be joined, producing a sequence of group memberships.

In the Mississippi Chat section, a plot of R-squared versus number of clusters shows a kink at the five-cluster level which might represent a natural subdivision of the data, since the preceding merger (from six clusters) produced relatively little increase in the R^2 value, and the next merger (to four clusters) produces a relatively greater increase. A display of partitions of the section ranging from two to five, shows a hierarchy that is consistent with the visual structure of the composition profile. The results can have a variety of applications, ranging from a layered representation of the section as input for a simulation model to a subdivision suitable for lateral stratigraphic correlation of either lithofacies or flow units. The implications of any depth-constrained clustering of logs can be targeted by the selection of the logs, so that their application can be made sensitive to mineralogical composition, pore volume, or other attributes.

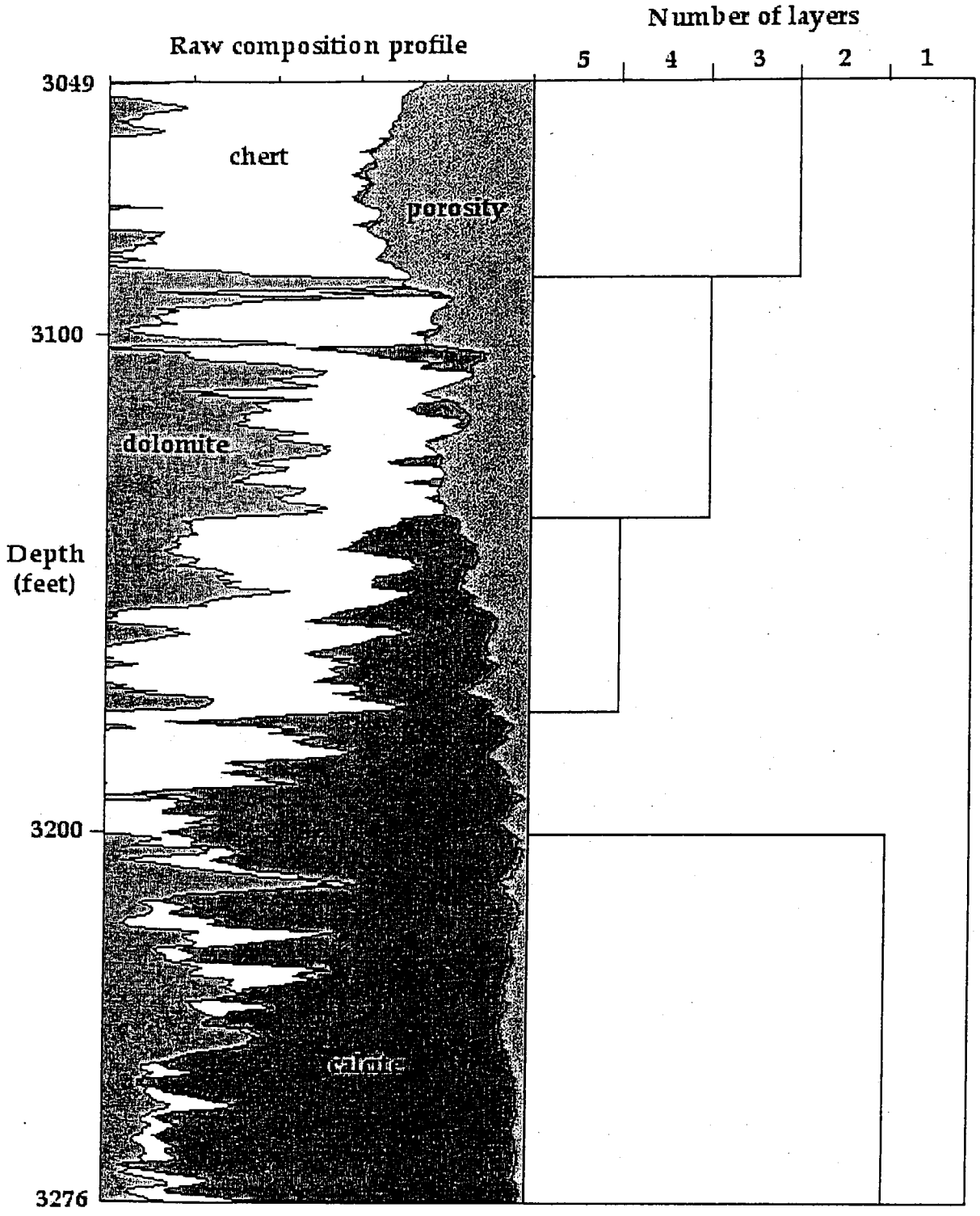


Plot of R-square versus number of zones in the Mississippian section of Pioneer Exploration Petrie #4 SE-NW-SW 36-26S-1W Sedgwick Co., Kansas



Pioneer Exploration Petrie #1 SE-NW-SW 36-26S-1W Sedgwick Co., Kansas

Composition profile plot of Mississippian carbonate section computed from density, neutron, and photoelectric factor curves (left) and zoned by five depth-constrained clusters (right).



Pioneer Exploration Petrie #1 SE-NW-SW 36-26S-1W Sedgwick Co., Kansas

Composition profile plot of Mississippian carbonate section computed from density, neutron, and photoelectric factor curves (left) and subdivided by one through five depth-constrained zones (right).

TREND SURFACE ANALYSIS

The empirical method of simple mapping may be formalized in a statistical model through the use of trend surface analysis. The theory is drawn directly from regression, which partitions observational data between a regional systematic component and a spatially uncorrelated "random" element. The regional component is estimated by a regression of the mapped variable on polynomials of the geographic coordinates of the well control. The regional surface is fitted to the data in such a way that the sum of the squares of their deviations from the surface is the minimum possible (Fig. 6).

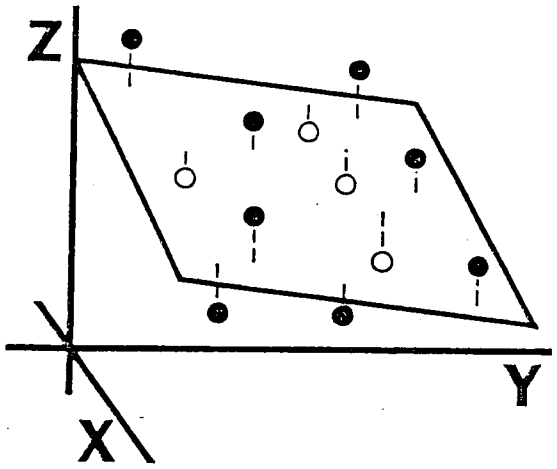


Figure 6: Hypothetical linear trend surface (a plane) fitted to a variable, Z, measured at locations with coordinates X and Y.

A hierarchy of trend surfaces matches polynomial equations of differing degrees of complexity. A linear (first order) surface is a plane; a quadratic (second order) surface is a paraboloid which may describe a basin or a positive feature. Higher order trend equations will represent successively more complex surfaces. As a regression procedure, trend surface analysis functions as an analysis of variance, in which the total variation of the data is partitioned between variation accounted for by the surface and the remainder as the sum of the squared deviations from the surface. When the residuals show no spatial correlation, the associated surface contains an estimate of the systematic spatial structure of the data variation.

This model can be applied to the mapping of a log response in a "calibration unit." A calibration unit is defined as a geological horizon which is judged to show only minor variation across the region of interest. The definition implies that a log response measured at any location can be subdivided into two components: a systematic regional element, and a random error associated with the miscalibration of the tool. The theory is best illustrated in a practical example of the method described by Doveton and Bornemann (1981).

A neutron log of the Viola limestone is shown from a well in south central Kansas (Fig. 7). At the base of the Viola is a distinctive "low porosity zone," which can be correlated in all wells of the area. The Lower limestone is essentially a pure limestone with only minor amounts of chert and dolomite, and appears to have a very restricted range of porosity on the order of three percent. Its simple mineralogy, moderately uniform porosity, and absence of hydrocarbons suggest that it is a reasonable choice for a log calibration unit.

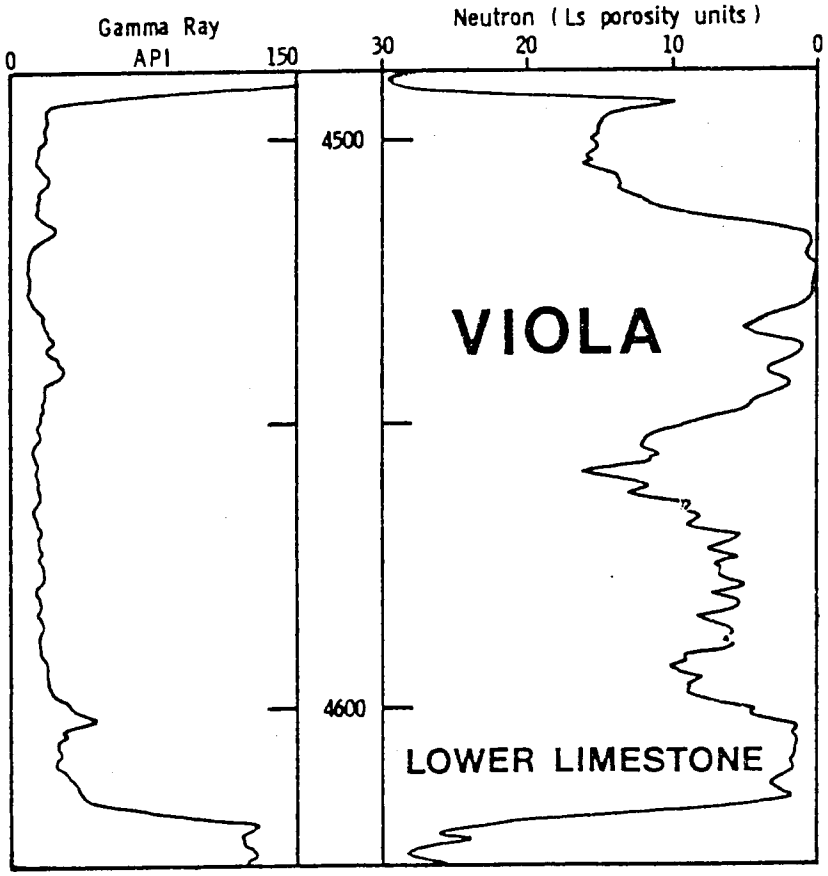


Figure 7: Representative neutron log of the Viola Limestone from a well in south-central Kansas.

Analytical data were drawn from 254 wells which penetrated the Viola and were logged by at least one porosity tool. The porosity logs were digitized over the range of the Viola and the total sample consisted of 194 neutron, 62 sonic, and 36 density logs. In each well, average neutron porosity, transit time, and density readings were calculated for the Lower Limestone, with an exclusion of the upper and lower two feet of the unit to minimize the effect of adjacent beds. The three sets of neutron, sonic, and density readings were made independent subjects for trend surface analysis.

Linear, quadratic, and cubic trend surfaces calculated for the Lower Limestone neutron data had fits of 6.72, 12.01, and 13.49 percent of the total variation. At first glance, these figures appear to be depressingly low, but they are an indication that the Lower Limestone is a good choice as a calibration unit. If the fits were high, the trend surfaces would imply that there was a major drift in the calibration standard across the area.

In moving from a linear to a quadratic surface, the degree of fit almost doubles, but the cubic is only a minor improvement on the quadratic. The quadratic surface appears to be the most "natural" expression of the systematic regional variation of the Lower Limestone neutron response. this interpretation was checked by a simple analysis of variance procedure which tests whether successively higher orders of surface make a statistically significant improvement over lower orders (Table 1). The linear surface does a significantly better job of predicting the true neutron response of the Lower Limestone at any location, when contrasted with use of the average neutron response in the data set. This demonstrates that the trend surface analysis has detected a systematic geographic component of variation in the calibration unit. The quadratic surface makes a significant improvement in fit over the linear, but the cubic fails to add significantly to the quadratic.

TABLE 1: Analysis-of-variance table of trend surfaces fit to the lower limestone neutron data. Asterisks indicate F-ratios significant at the 95 percent level.

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	21.41	3	7.14	3.13*
Quadratic deviation	356.22	156	2.28	
Cubic-quadratic	5.99	4	1.50	0.65
Cubic deviation	350.23	152	2.30	
Total variation	404.85	161		

The configuration of the quadratic surface (Fig. 8) shows a broad central area of relatively high porosity flanked by lower porosities to the west and southeast. It is worthwhile to relate log response trend surfaces with regional geology patterns in order to assign meaning to the trends and to cross-check their validity. The quadratic surface shows a striking concordance with the axis of the Pratt anticline and suggests a genetic relationship. The neutron trend surface indicates that the Pratt anticline was an active positive feature contemporaneous with the deposition of the Lower Limestone. The trend in porosity may well reflect a regional pattern of the low porosity, grain-support crinoidal packstones and grainstones on the margins of the Pratt anticline, which grade towards the axis as a slightly more micritic and higher porosity facies equivalent.

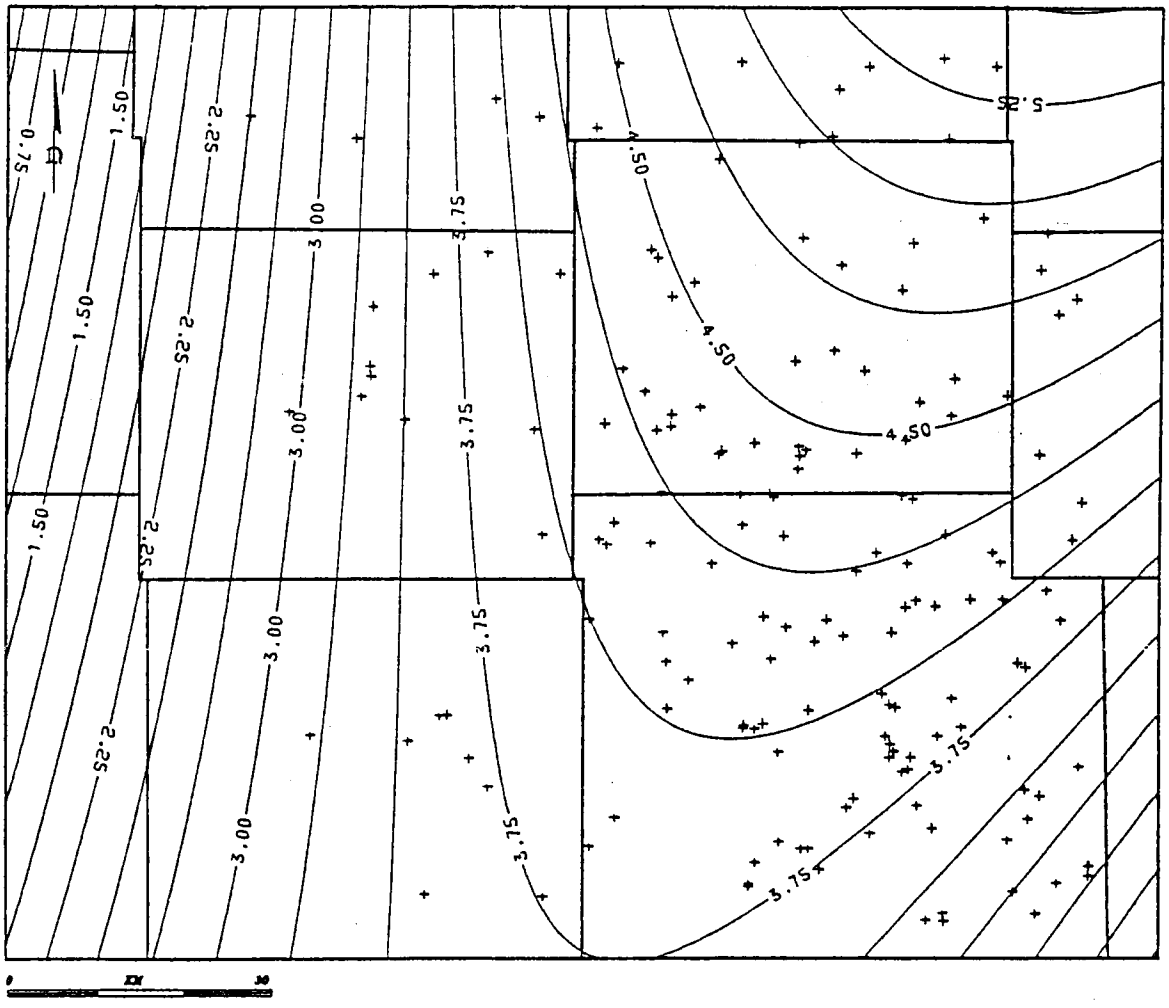


Figure 8: Quadratic trend surface map of neutron porosity variation of the Viola "Lower Limestone" zone in south-central Kansas.

Trend surfaces were computed for the transit times and densities of the Lower Limestone and, in both cases, the linear surface was found to be most significant in portraying the regional variation of the data. A contour map of the transit time linear surface (Fig. 9) shows a simple decline in porosity, moving from the northeast to the southwest, which is only a general approximation of the trend in the neutron data. A major reason for this difference is that the sonic log sample is less than one-third the size of the neutron sample. As a result, the restricted sample size is not large enough to pick up any small systematic improvement by the quadratic surface.

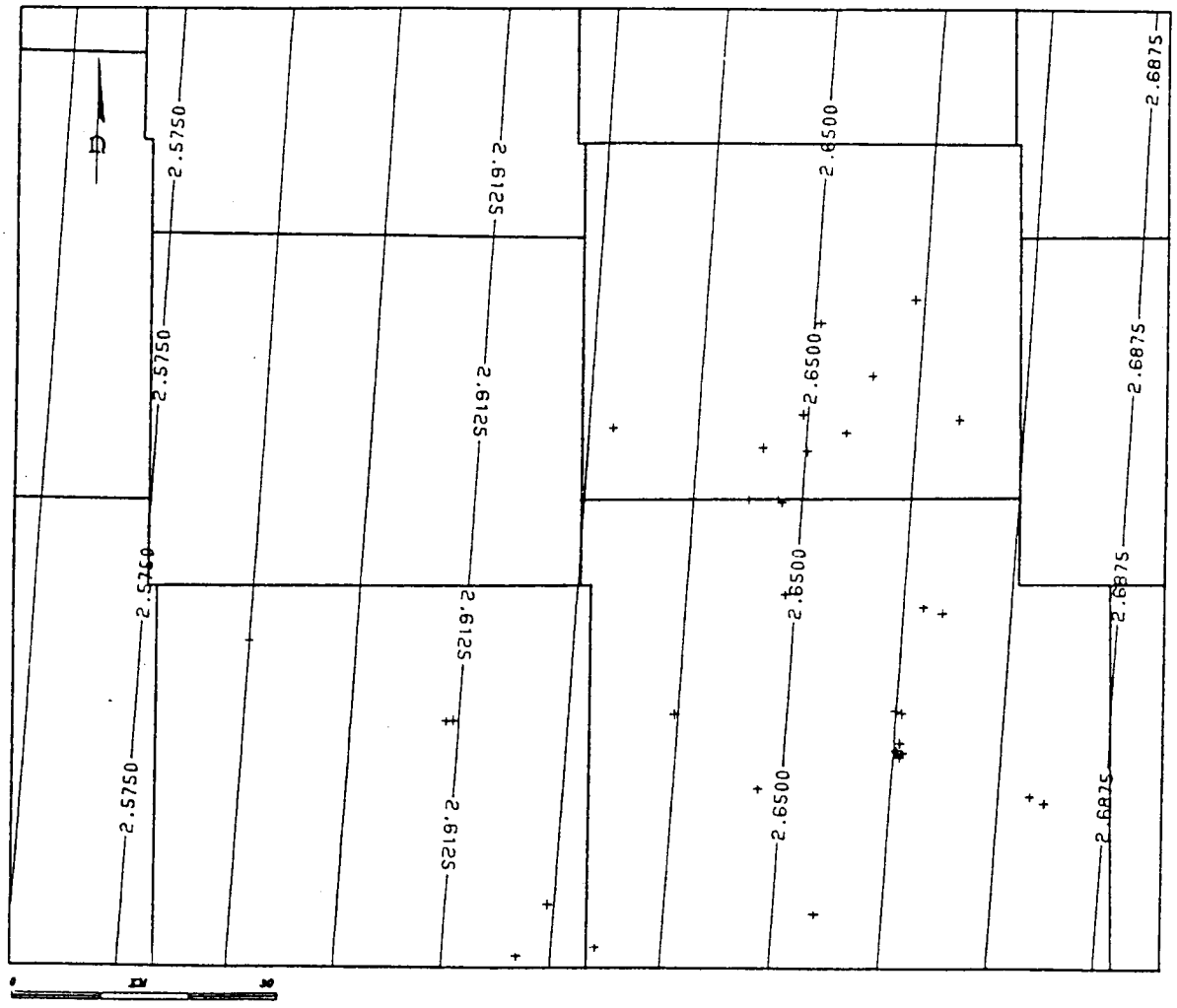


Figure 9: Linear trend surface of Lower Limestone bulk density variation (grams per cubic centimeter)

The same argument applies to the density linear surface where the sample is further reduced to 36 wells. However, the density linear surface (Fig. 10) shows a decline in porosity to the east which seems to contradict the trends observed in both neutron and sonic surfaces. The apparent discrepancy is explained by the distribution of the density well control, which is almost entirely restricted to the eastern half of the area. The density surface is therefore primarily sensitive to the fall in porosity on the eastern flank of the regional structure. This characteristic makes the important point that trend surface predictions of calibration unit response must be restricted to areas of moderate well control and not extrapolated beyond their limits.

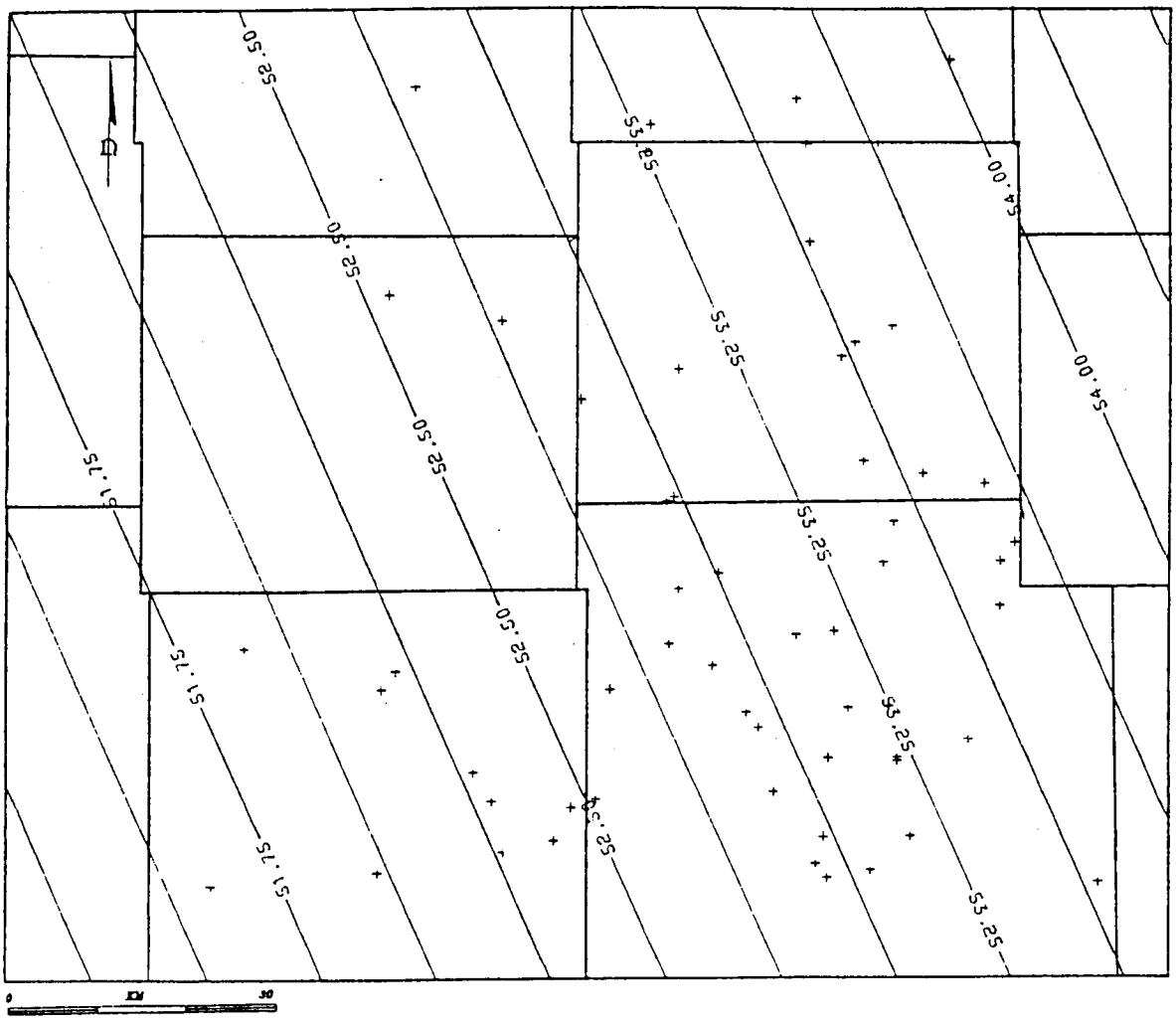


Figure 10: Linear trend surface of Lower Limestone transit time variation (microseconds per foot).

While the trend surfaces of the Lower Limestone porosity values are estimates of the systematic regional variation, the deviations of the raw values from the surface are attributed to a contribution of tool 11 error and any systematic local variation. If the residuals are controlled by tool error alone, they should be approximately normally distributed with a mode located at zero. Frequency polygons of the neutron, sonic, and density residuals are shown in Figure 11, plotted on a compatible scale of limestone porosity units. In all cases, the distributions show a slight positive skew with a displacement of the modal peak to approximately minus one porosity units. This common feature implies that the true variation of the Lower Limestone closely follows the form of the trend surfaces, but is broken in local areas by the development of either enhanced porosity zones or increased shale content. As a result, the trend surface estimates will tend to be slightly optimistic over most of the area by about one porosity unit, but locally pessimistic by approximately two porosity units. Following conventional practice, each trend surface equation can be modified by subtracting the displacement of the mode from zero and inserting this quantity as a correction factor. The modified equation is now geared to the most typical values for the area, rather than the grand average. Potential problem areas of enhanced porosity can be examined through contour maps of the residuals, although it will often be difficult to distinguish spurious features caused by positive tool errors.

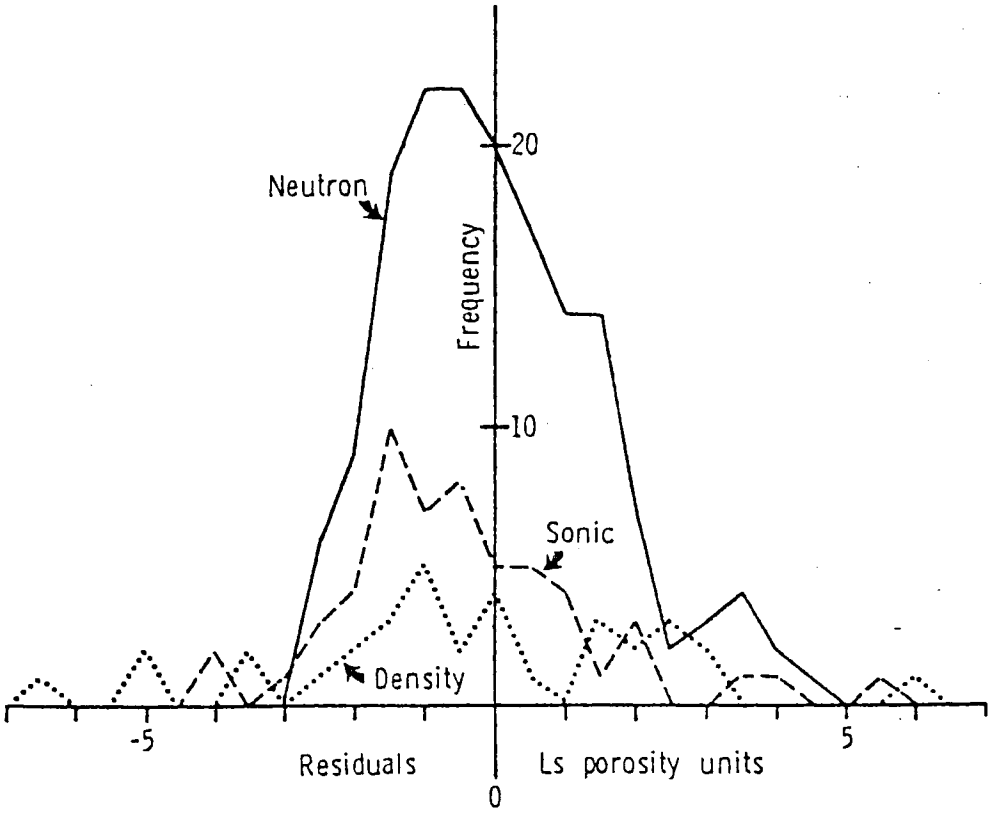


Figure 11: Frequency polygons of Lower Limestone neutron, sonic, and density trend surface residuals.

The analysis of both the trend surfaces and their associated residuals therefore provides useful feedback concerning the performance of the Lower Limestone as satisfactory calibration unit, as well as numerical relationships for prediction and control. The trend surface equations are used in prognosis, while the residual distributions are useful for diagnosis. Simple descriptive statistics and inferential tests may be applied in a systematic dissection of log quality in the area if the residuals can be reasonably approximated by normal distributions. The area under the curve of a normal distribution is a function of a scale of standard deviations about the mean and dictates the proportion of tool errors to be expected within specified ranges. So, for example, if an accuracy of plus or minus one porosity unit is considered an acceptable tolerance level, 49 percent of the neutron logs will require correction, since the standard deviation of the neutron is 1.51 porosity units. This figure compares with 69 percent of both sonic and density logs.

The trend surface normalization method can be expanded to a generalized strategy of analysis designed for routine applications. The major steps are:

(1) Define an area of interest in which well logs are to be normalized.

(2) Select zones from the regional stratigraphy which are both easily correlated and appear to show only minor changes in logging character across the area. These zones are provisional calibration units whose suitability as normalization standards is to be examined in the first phase of analysis.

(3) Compile data sets for each provisional unit to include geographic coordinates and log responses. For most logs, the raw response is appropriate. However, resistivity values should be transformed to the square root of the conductivity in order to approximately linearize the variation. Any necessary environmental corrections should be applied.

(4) For each provisional unit, compute trend surfaces for the separate log responses. Extract a specific surface from each trend surface series which appears to match the regional variation, as judged from inspection of the degrees of fit or by analysis of variance. Compare maps of the selected trend surfaces with regional geological expectations of depositional facies variations and compactional trends as an external check on their validity. The residuals of an ideal calibration unit should be approximately normally distributed with a mode located at zero. Displacement of the mode and skewness of the residual distribution or the development of distinctive secondary modes all indicate localized areas of systematic variation which has been compounded with tool error. However, some allowance for deviations from the ideal should be made when dealing with small sample sizes. Provisional units with satisfactory trend surface and residual properties may be adopted as calibration units for prediction and control.

(5) Compute prognoses of calibration unit normalized responses at existing well control or prospective drilling sites by insertion of the geographic coordinates into the trend surface equations. In practice, it is wise to run several

calibration units as a crosscheck on consistent normalization corrections and to highlight possible changes in tool error as a function of depth. Prognosis should not be made outside the area of well control, and the overall reliability of any prognosis will be controlled by the density of neighboring wells.

(6) For more detailed studies, analyze the residual distributions in assessments of error character contrasted between tools and between service companies in terms of relative precision and bias.

(7) Update calibration unit data sets with log responses and geographic coordinates from newer control as it becomes available. Periodically, recompute trend surfaces as successively more refined estimates of regional variation.

The procedure outlined is extremely straightforward as a practical program for log normalization. Programs for trend surface analysis are widely available since they are used by exploration geologists in structural studies. These programs are cheap to run and their output routinely lists trend surface equations, residuals, and analysis of variance data, as well as generating simple maps or files for the plotter production of maps.

Example: Use of trend surface analysis in the search for Upper Devonian reefs in Alberta, Canada

The stratigraphic succession overlying Upper Devonian Leduc reefs shows a pronounced structural drape whose magnitude diminishes upward, but is still distinctive in horizons as high as the Upper Cretaceous. The geological structure of southwest Alberta is dominated by a strong regional dip towards the Rocky Mountain chain, so that local structural drape over Leduc reefs is manifested as minor perturbations to the main structural grain. As a result, it is difficult to distinguish local structural highs from the regional trend in raw structural surfaces. The structure of any horizon in this area may be considered to be the sum of three components:

(1) trend (the regional tectonic surface);

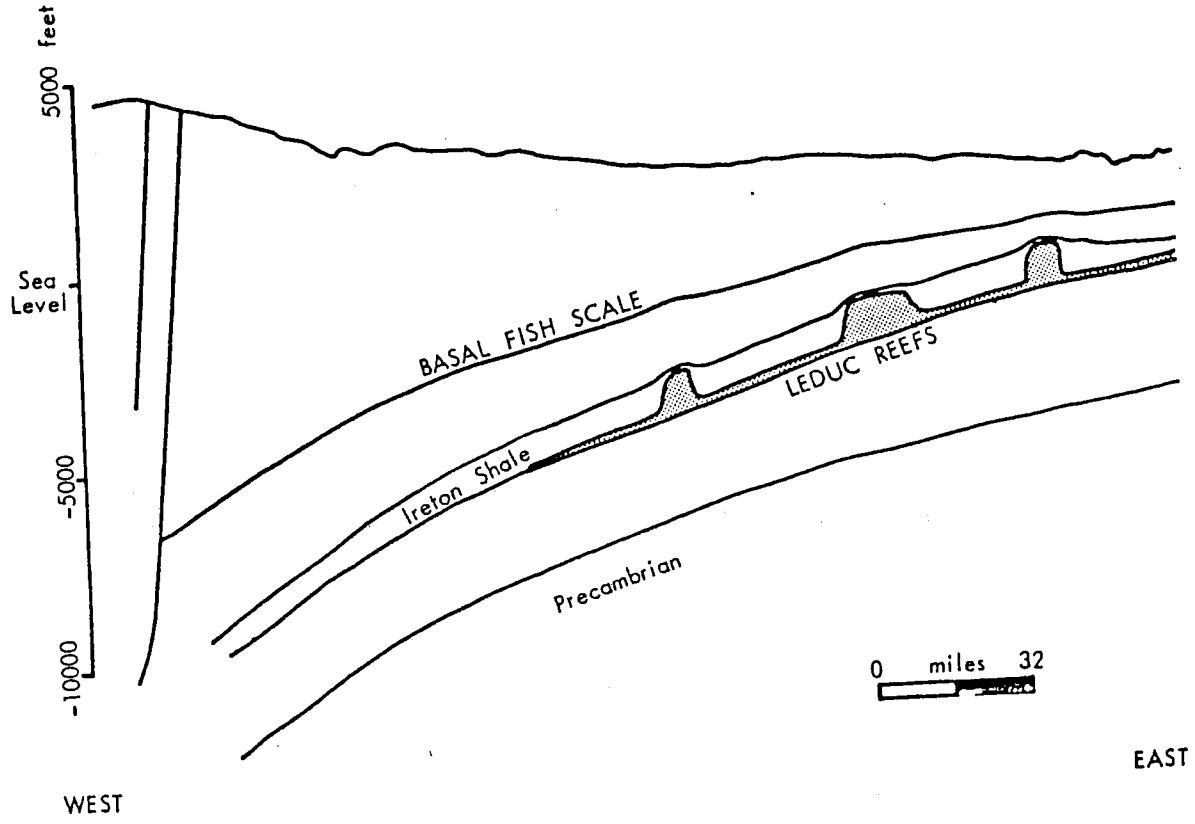
(2) local anomaly (structural drape over reefs);

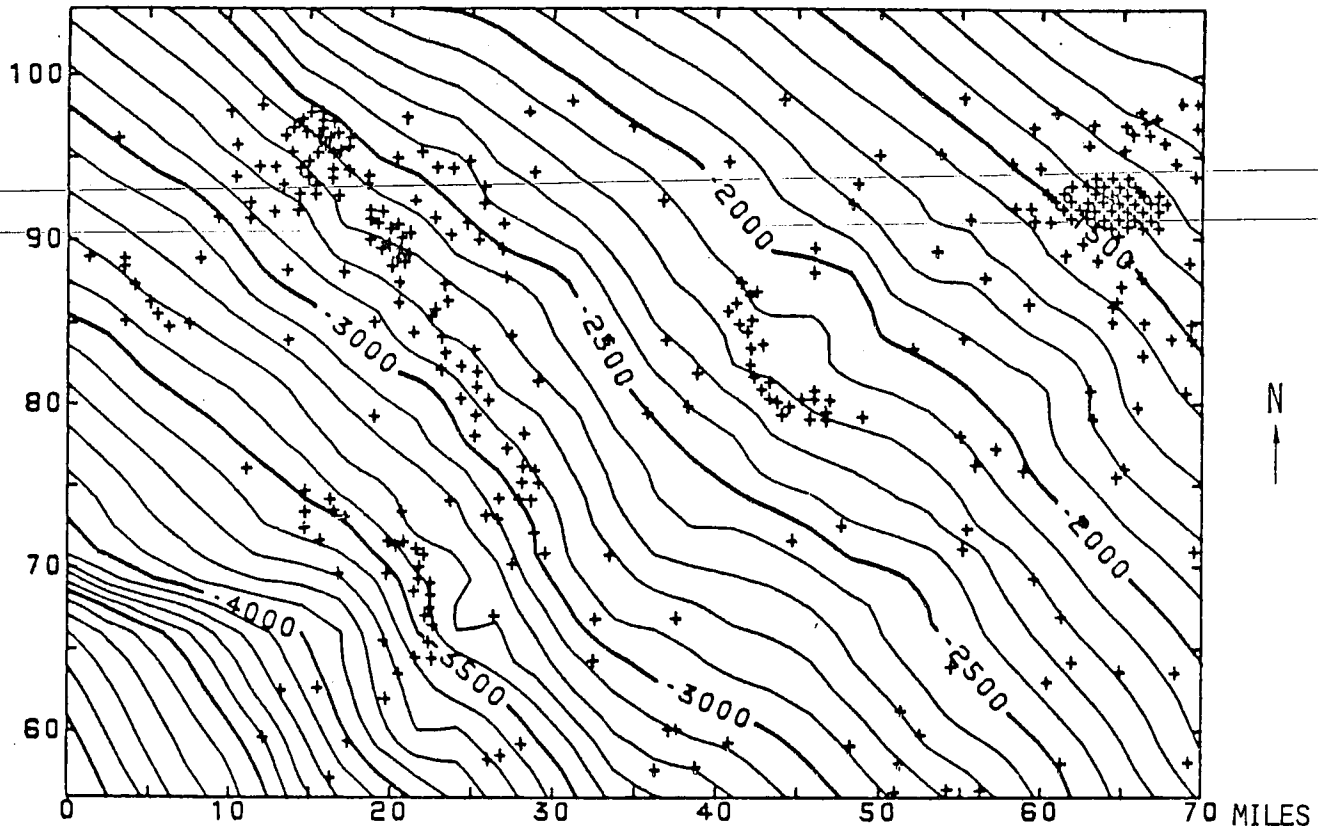
(3) error (mis-picks of the horizon, miscalibration of logs, etc.).

The base of the Fish Scales Sandstone (Lower Cretaceous) is the most commonly used horizon in searching for structural drape anomalies caused by underlying Devonian reefs. This horizon is selected for two reasons. Firstly, the base of the Fish Scales is highly radioactive, appearing as an almost unmistakeable spike on gamma ray logs, so that the chances of a mis-pick (error) are reduced to a minimum. Secondly, since the horizon is relatively shallow, there is a significantly greater degree of well control than is available for the Devonian. As a result, local structural anomalies may be mapped in areas where there are few or no wells that penetrate the Devonian and the method used as an exploration tool.

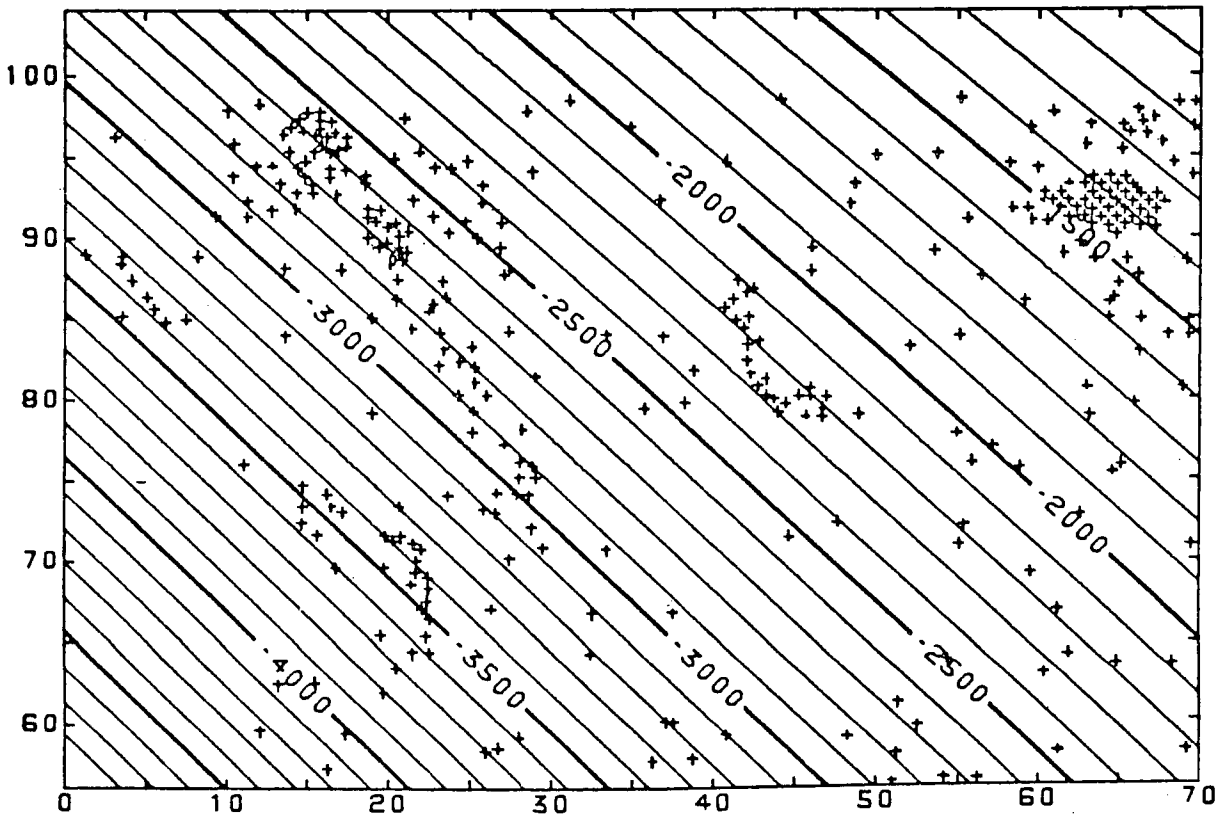
The illustrated maps show the relationship between processed basal Fish Scales structure and the distribution of Leduc reefs in the Windfall Reef area, south-central Alberta. The raw structural map is dominated by the regional dip of the horizon to the southwest and reef-related anomalies are difficult to perceive. A linear trend surface fitted to the horizon has a fit of 98.6% and may be equated with the regional component. However, the regional structural surface shows a small but systematic curvature, with increasing dip to the southwest, as is shown by the increasing constriction of the contour lines in that direction.

Consequently, a second-order surface (corresponding to a parabolloid) is more truly representative of the regional trend and provides a fit of 99.7%. By subtracting the second-order surface from the raw structure, second-order residuals may be mapped which are composed of local structural anomalies together with any error. The correspondence between positive residual anomalies and the Windfall Reef complex (as deduced from available Devonian well information) is very striking. Devonian tests in the shelf area to the east are relatively few and positive residual anomaly areas in this region are worth close examination as possible fringing reef prospects.

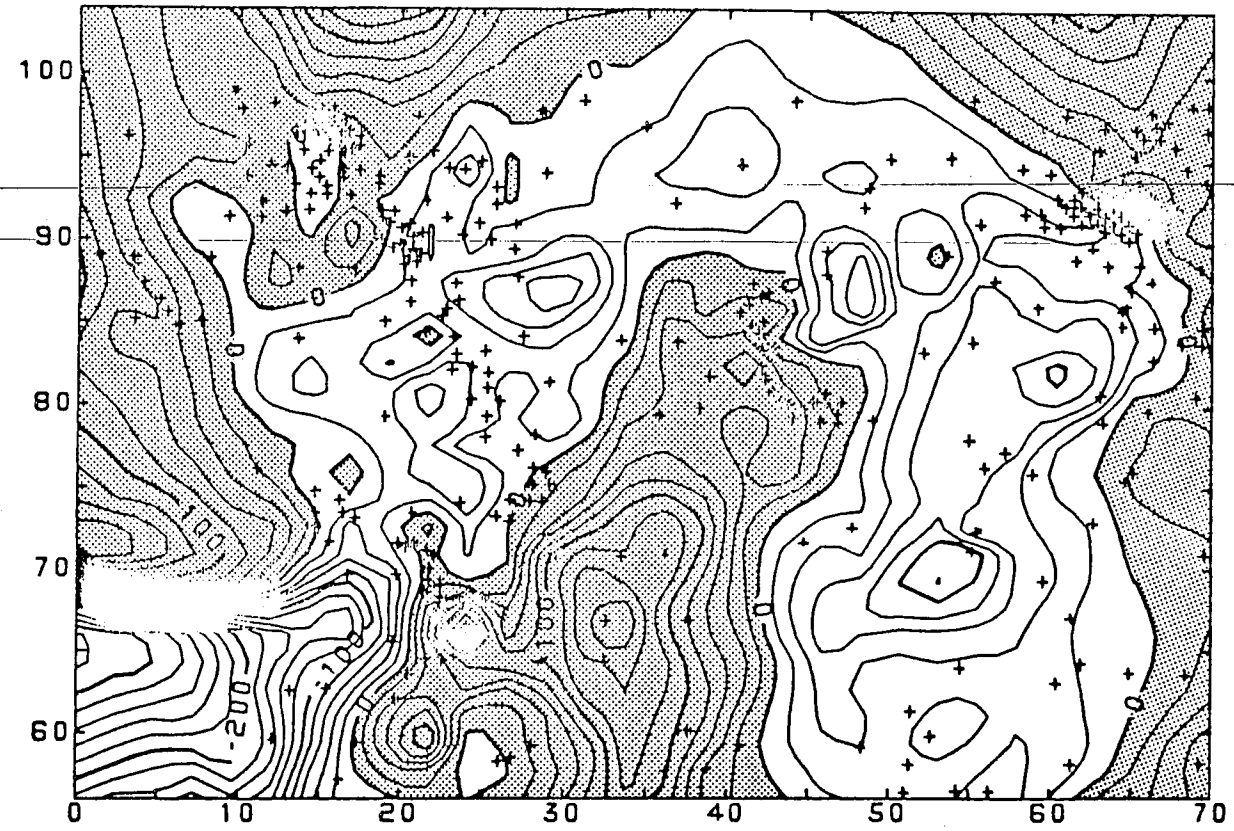




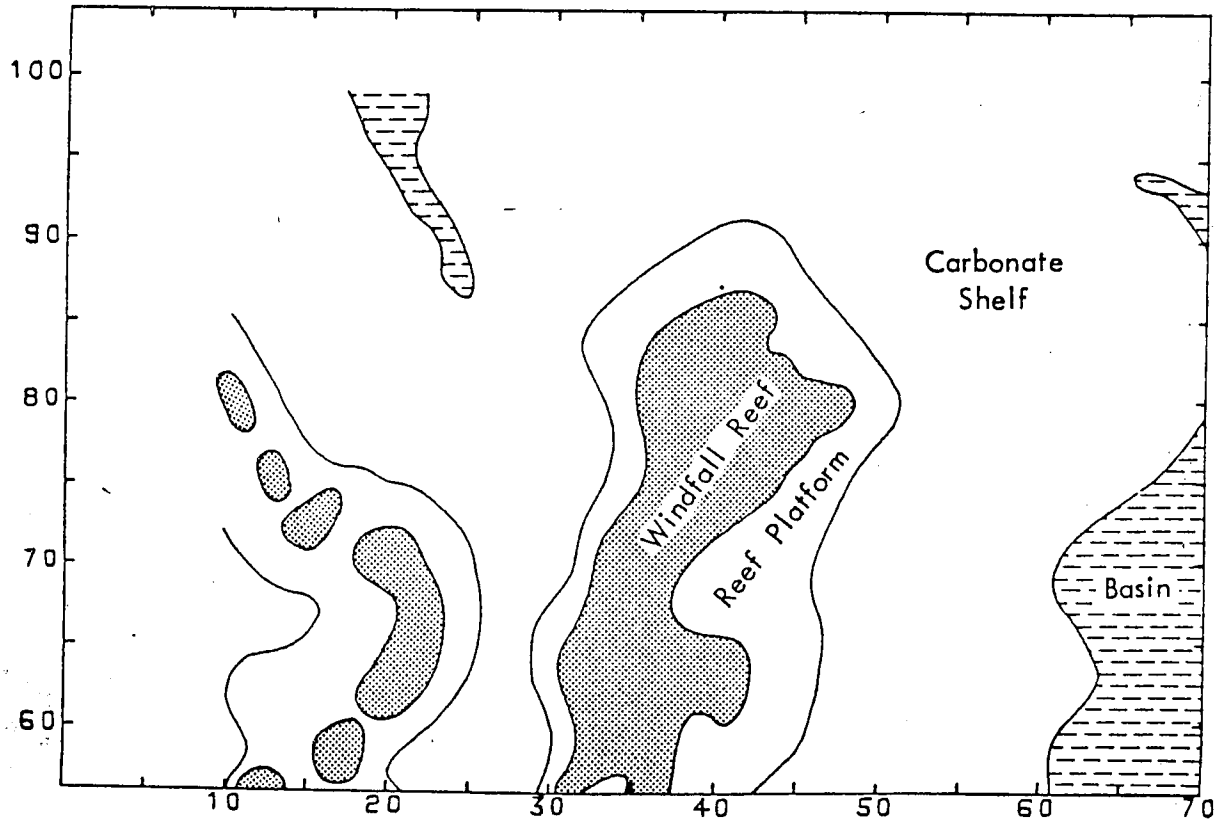
Basal Fish Scales structural surface



Second-order trend surface fitted to basal Fish Scales



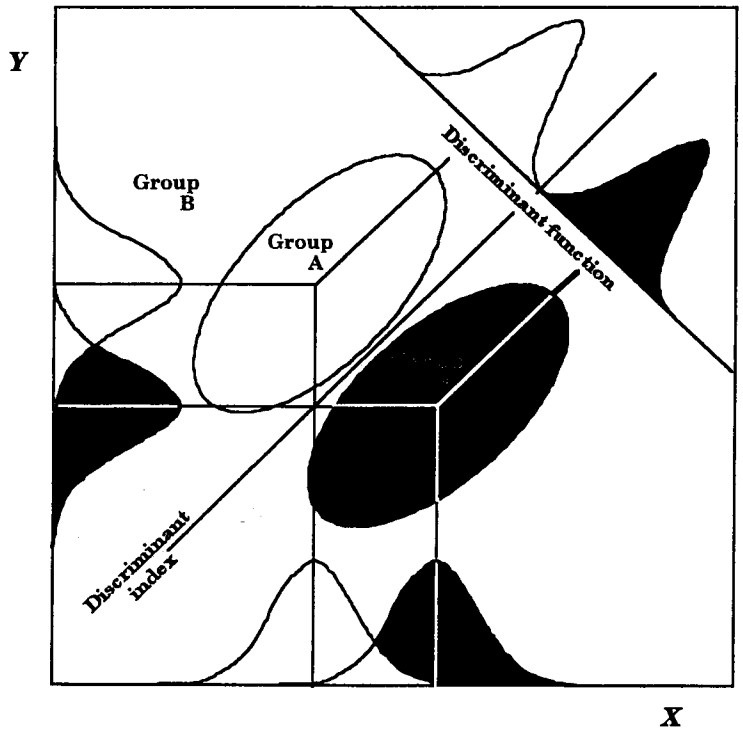
Basal Fish Scales second-order trend surface residuals



Leduc reef paleogeography as deduced from Devonian wells

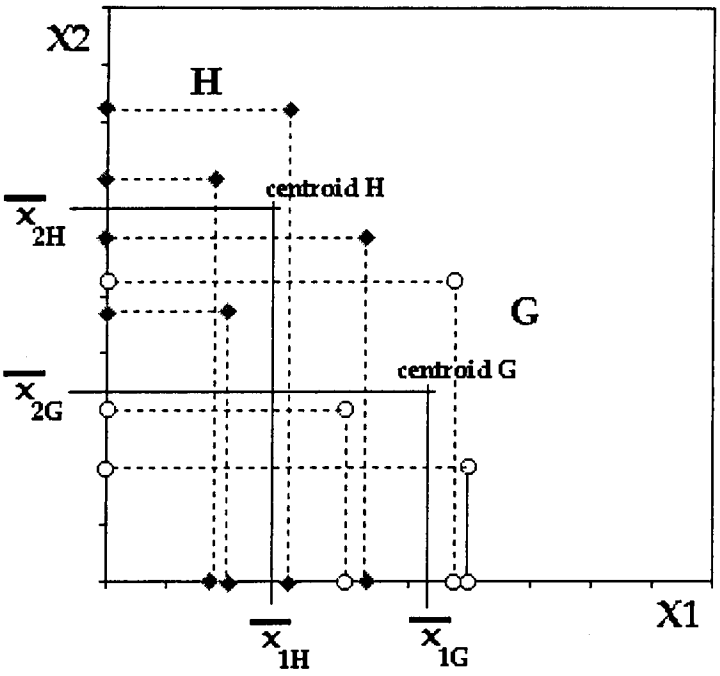
DISCRIMINANT FUNCTION ANALYSIS

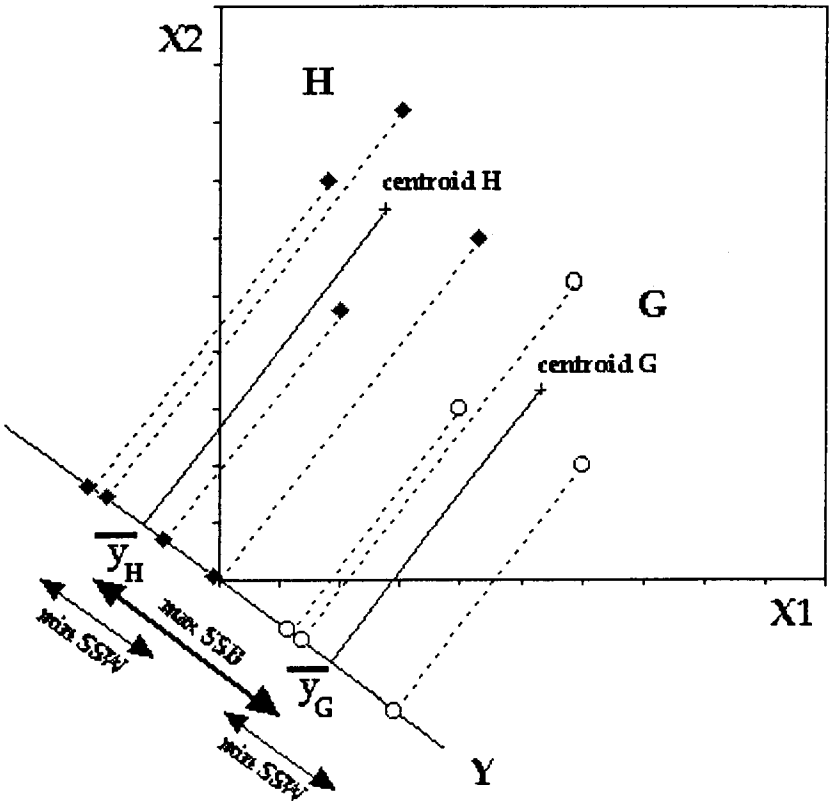
Discriminant analysis covers a wide range of techniques aimed at the classification of unknown samples to one of several possible groups or classes. Classical discriminant function analysis has the tighter focus of attempting to develop a linear equation that best differentiates between two different classes. Fisher (1936) first derived the linear discriminant function as a statistical method to separate two populations by a linear weighted function of their measurement variables. The method is a supervised technique that requires a training data set for which assignations to the two populations are already known. The data consist of multivariate values for every individual in both population samples. If the two groups were plotted in multidimensional space, they would appear as two clouds of data points with either a distinctive separation or some degree of overlap. An axis is located on which the distance between each cloud is maximized while the dispersion within each cloud is simultaneously minimized. This axis defines the linear discriminant function and is calculated from the multivariate means, variances, and covariances of the two groups. The data points of the two groups may be projected onto this axis as locations on a single line. This operation results in the collapse of the many variable dimensions of the recorded data into a single, composite variable that best discriminates between the two groups.



The process is perhaps best understood by reference to the diagram of a hypothetical small problem in two-dimensions. When two groups are plotted with respect to a single variable there is frequently a range of overlap, as shown by the frequency curves on each axis. When an observation of unknown affinity occurs in this range there is a degree of uncertainty in its classification. It can be assigned to one of the groups by a decision based on probability, but the assignment will be incorrect in (hopefully) a minority of occasions. When plotted in the Cartesian space of two variables, the degree of differentiation will

improve or, at worst, stay the same. An ideal situation is shown with a perfect separation between the two groups. In practice, there will commonly be an area of intermingling of the two data-point clouds. The discriminant function is then the equation of the axis that cuts obliquely across the crossplot plane at an angle determined by the relative contribution of the two variables to discrimination. In the computation of the function, the two clouds are modelled by bivariate normal distributions whose probability density contours map out elliptical shapes. When a higher number of variables is used for discrimination the two clouds are matched with hyperellipsoids in multivariate space.





The equation of the discriminant function, y, is:

$$y = w_1x_1 + w_2x_2 + \ldots\ldots w_mx_m$$

where x_1 to x_m are the *m* variables used for discrimination and w_i is the weighting coefficient to be applied to the variable x_i. The optimum discriminant function is set at that location which maximizes the value of the distance between the cloud centroids divided by the dispersion of the two clouds. This condition occurs when the following equation is satisfied:

$$\mathbf{SW = D}$$

where **S** is a matrix of pooled variances and covariances from the two groups, **W** is a vector of the unknown weights to be solved, and **D** is a vector of the differences between the means of the two groups with respect to the variables.

There is a simplifying assumption that the variance-covariance matrix is the same in the two groups. So, the pooled variance-covariance matrix is an average for the two groups weighted by their number of observations. This assumption implies that the two group data clouds have similar inflations and orientations. Whether the stipulation is reasonable can be checked by a simple visual inspection of crossplots. The method is sufficiently robust that moderate differences do not generally degrade the power of the function as an effective classifier. Indeed, as Reyment (1974) pointed out, the original example used by Fisher (1936) clearly violates this assumed equality of the dispersion matrices. Heterogeneous discriminant functions can be applied for particularly troublesome classification problems. However, the simple linear discriminant is usually quite adequate for most applications.

The solution of the matrix equation for the vector of discriminant coefficients then follows as:

$$\mathbf{W = S^{-1} D}$$

The function describes a linear axis that crosses the multivariate space. The coordinates of any individual zone can be converted into a discriminant score, y_i, which is a projection onto the discriminant axis and gives its relative position:

$$y_i = w_1x_{1i} + w_2x_{2i} + \ldots.. w_mx_{mi}$$

where x_{1i} to x_{mi} are the values of the *m* variables in the i^{th} zone. The multivariate data clouds are now condensed into two ranges of points on a single axis. A discriminant index is calculated as the midpoint between the discriminant score projections of the two group multivariate means. The discriminant index serves as a boundary between the domains of the two groups and can be used as a decision criterion to identify an unknown observation with one or the other of the groups. The choice of midpoint reflects the assumption of equal dispersion matrices for the two groups. However, the discriminant boundary can be adjusted to a better value, based on the rate of misclassifications observed in validation tests of the function.

The mathematics of the discriminant function will generate some form of solution, even when there are no systematic differences between the two groups. This contingency can be checked statistically by an F-test, based on a generalized measure known as Mahalanobis' distance, . This represents the relative degree of separation between the centroids of the two groups, given by:

$$D^2 = w_1d_1 + w_2d_2 + \ldots\ldots + w_md_m$$

where d_i is the difference between the means of the two groups for the i^{th} variable.

Mahalanobis' distance can also be used to assess the relative contribution of each measurement variable to discrimination. This is computed through a comparison of the weighted difference of the group means:

$$E_i = \frac{w_i d_i}{D^2}$$

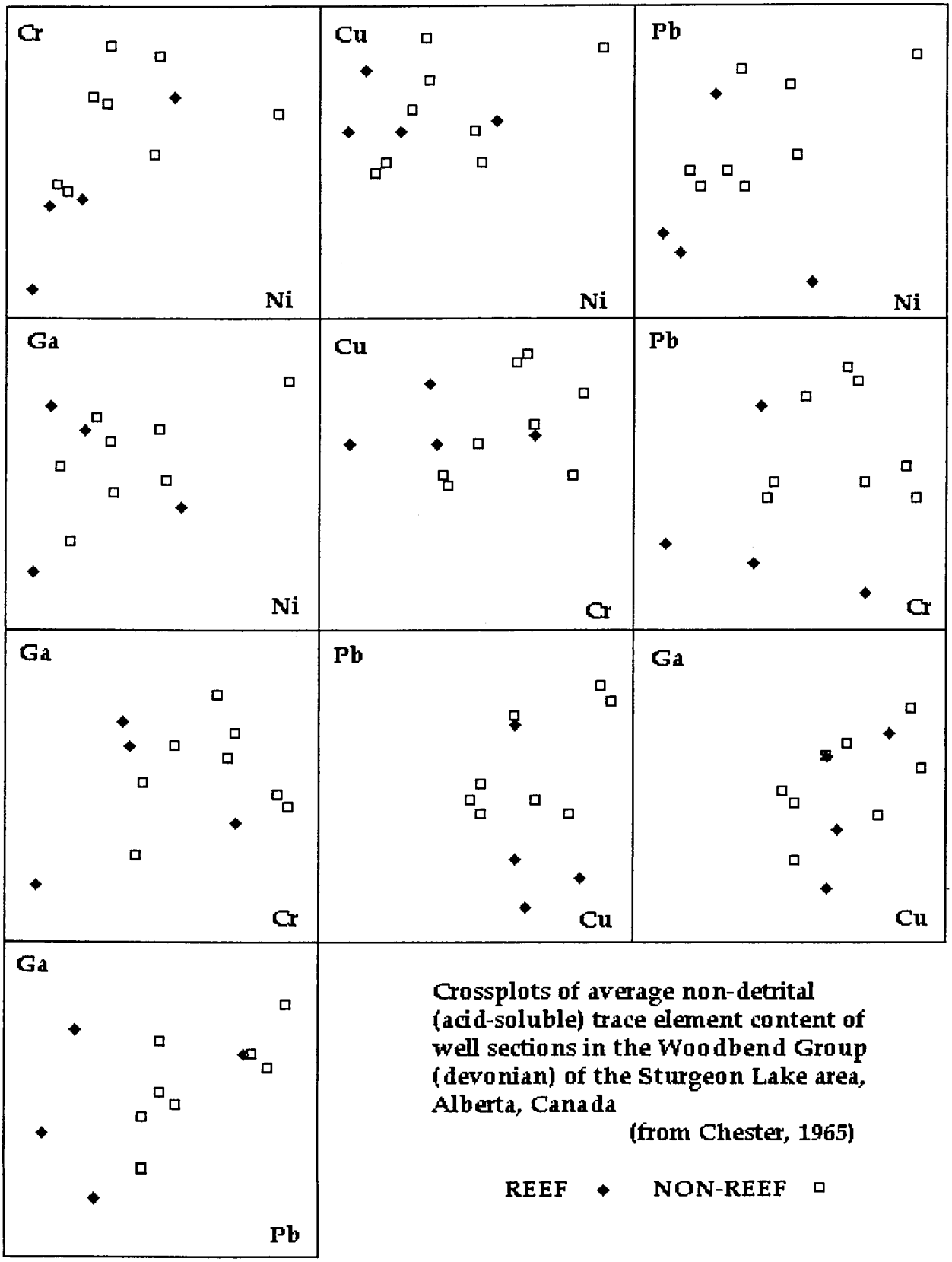
where E_i is the contribution of the i^{th} variable. In addition, the distance function can be used to provide a means to judge whether an unknown observation belongs to one of the two groups or not. In this case, the distance is computed between the multivariate location of the zone and the centroid of each group. Because the distance is computed in multivariate space, it is useful as an error diagnostic to alert the user to situations where a zone is likely to belong to a third group.

Finally, a table of correct and incorrect classifications for the calibration data set and in addition, preferably, a validation set, will also give a direct measure of classification performance. The results of such tables can be used to evaluate overall success in predicting classifications, compare different models of log measurements, and monitor possible improvements in prediction made through corrective adjustments of the discriminant index.

DIFFERENTIATION OF TRACE ELEMENTS IN DEVONIAN REEF AND NON-REEF CARBONATES: A PETROLEUM EXPLORATION CASE STUDY

Exploration holes that penetrate the Upper Devonian Woodbend Group in Alberta, Canada are either successful in locating a Leduc reef or fail by drilling through a section of the laterally equivalent Ireton Shale. Typically, they are drilled on seismic prospects that indicate a reef location, so that a failure could be a near miss. If this is so, then a successful well could be completed by either directional drilling or twinning the well. Alternatively, the exploration hole could be far removed from a reef and so should be abandoned as a dry hole. Off-reef sections frequently have carbonates interbedded with shales, so trace element geochemical analysis was made of the acid-soluble content of carbonates from Leduc reefs and contrasted with carbonates in the Ireton Shale that represent basinal facies. If a satisfactory distinction could be made, then the results could be used in the assessment of carbonates from a potential near miss as to whether they were likely to be from the talus of as nearby reef or from a basinal facies carbonate. Crossplots are shown of average trace element compositions from carbonates in reef and non-reef wells in the Sturgeon Lake area. The data are summarized from more extensive analyses published by Chester (1965).

	ID	Ni	Cr	Cu	Pb	Ga
REEF	24	23	3.6	18	29	16
	33	13	3.4	24	8.6	18
	46	8	0.9	18	11	4.7
	32	51	6.6	19	4.7	9.8
NON-REEF	11	82	6.1	26	34	20
	6	45	4.9	18	30	16
	13	19	3.8	15	17	7
	12	16	4	14	19	13
	51	47	7.8	15	21	12
	67	27	6.6	20	19	17
	61	31	6.4	27	32	15
	18	32	8.1	23	17	11



Crossplots of average non-detrital (acid-soluble) trace element content of well sections in the Woodbend Group (devonian) of the Sturgeon Lake area, Alberta, Canada

(from Chester, 1965)

REEF ◆ NON-REEF □

As a simple demonstration, we can show the results of discriminant analysis using just two elements - nickel and lead - to differentiate the reef and non-reef wells, before using all five elements.
The data are:

	ID	Ni	Pb	GROUP
REEF	24	23	29	1
	33	13	8.6	1
	46	8	11	1
	32	51	4.7	1
NON-REEF	11	82	34	2
	6	45	30	2
	13	19	17	2
	12	16	19	2
	51	47	21	2
	67	27	19	2
	61	31	32	2
	18	32	17	2

The group means are:

Group	Count	Ni mean	Pb mean
1	4	23.75	13.33
2	8	37.38	23.63
Global	12	32.83	20.19

The pooled covariance matrix for the groups (**S**) is:

	Ni	Pb
Ni	422.06	57.78
Pb	57.78	70.37

The inverse of the covariance matrix is $\mathbf{S^{-1}}$:

0.00267	-0.00219
-0.00219	0.01601

The vector of the differences of the means of the two groups, **D** is:

-13.625	-10.3

Therefore, the weights are $\mathbf{W=S^{-1}D}$ which results in:

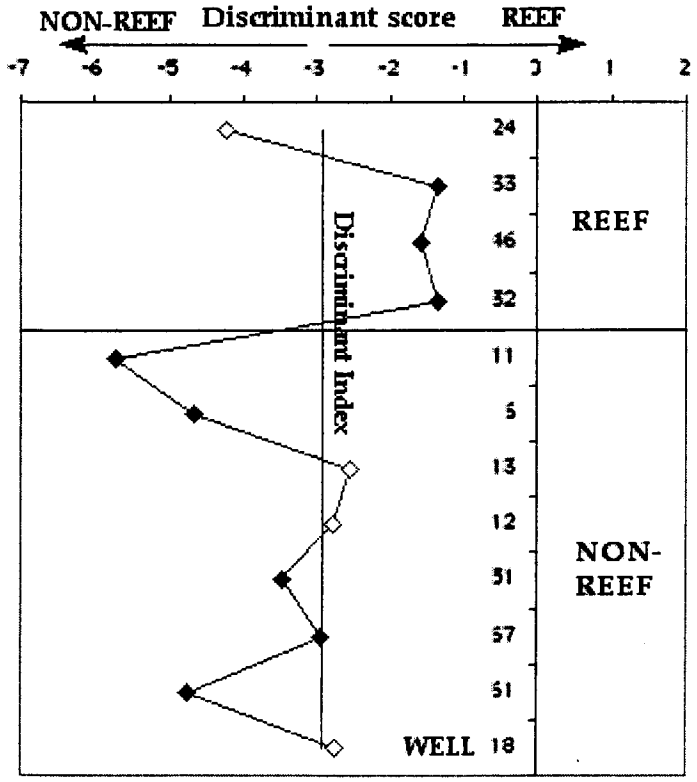
-0.0138	-0.1350

The discriminant scores are then calculated for each well by the equation:

$$y_i = w_1 x_{1i} + w_2 x_{2i} + \ldots . w_m x_{mi}$$

Group	ID	Ni	Pb	Score
1	24	23	29	-4.23
1	33	13	8.6	-1.34
1	46	8	11	-1.60
1	32	51	4.7	-1.34
2	11	82	34	-5.72
2	6	45	30	-4.67
2	13	19	17	-2.56
2	12	16	19	-2.79
2	51	47	21	-3.48
2	67	27	19	-2.94
2	61	31	32	-4.75
2	18	32	17	-2.74

The discriminant index is the discriminant score of the midpoint between the two group centroids and is -2.92. A plot of the well scores and the discriminant index shows one reef well and three non-reef wells misclassified.



Mahalanobis' distance is given by $D^2 = w_1 d_{Ni} + w_2 d_{Pb}$ which calculates as 1.579

The relative contribution to discrimination by Nickel is estimated by $E_{Ni} = \frac{w_1 d_{Ni}}{D^2}$ which equals 0.119 . The relative contribution to discrimination by Lead is estimated by $E_{Pb} = \frac{w_2 d_{Pb}}{D^2}$ which equals 0.881. So, using Nickel and Lead alone, Lead provides 88% of the discrimination.

Discrimination between reef and non-reef wells using all five trace elements gives the following results.
The mean values of the groups are:

Group	Count	Ni	Cr	Cu	Pb	Ga
1	4	23.75	3.63	19.75	13.33	12.13
2	8	37.38	5.96	19.75	23.63	13.88
Global	12	32.83	5.18	19.75	20.19	13.29

The pooled covariance matrix for the groups is:

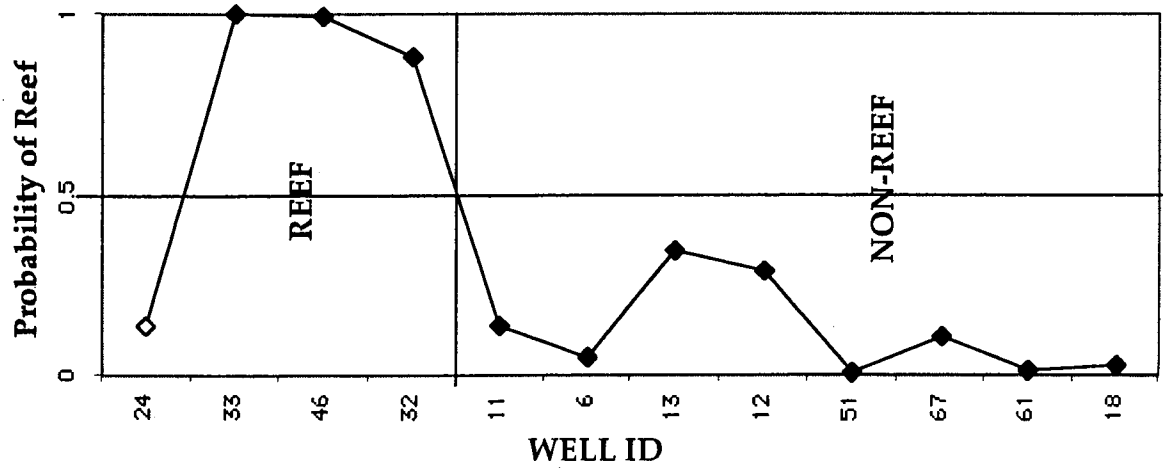
	Ni	Cr	Cu	Pb	Ga
Ni	422.06	20.29	32.65	57.78	37.99
Cr	20.29	3.45	2.75	-1.89	2.00
Cu	32.65	2.75	20.83	12.53	11.37
Pb	57.78	-1.89	12.53	70.37	21.59
Ga	37.99	2.00	11.37	21.59	22.29

The solution for the weights of the linear discriminant function are solved to be:

Ni	Cr	Cu	Pb	Ga
0.0447	-1.3602	0.2069	-0.3043	0.1564

Estimation of the relative discriminatory power of the trace elements indicates that chromium and lead dominate the distinction between reef and non-reef carbonates.

Although the score values give a sense of which group any well should be assigned to, predictions are better made in terms of probability. The calculation requires a prior probability to be assigned within a Bayesian equation that incorporates the discrimination statistics in the estimation of the posterior probability that the well belongs to one or other of the groups. Typically, the prior probability will be set as either equal an probability for each group or probabilities equal to the proportions of observations in each group. The calculation then assumes that the observations are multinormally distributed about the centroids of their groups. Assuming an equal probability of occurring in either reef or non-reef as the priors, the posterior probabilities for the wells, based on the five trace elements is:



Of course, the ultimate purpose of using discriminant functions to see if reef carbonates could be distinguished from non-reef carbonates was to see whether a classification method could be developed so that in a reef prospect well that did not penetrate a reef, an evaluation could be made from trace elements measured from the acid-soluble fraction of carbonates in drill-cuttings as to their likelihood of proximity of the borehole to a nearby reef. In one such well, the trace element composition was analyzed as:

Nickel 32 Chromium 4.2 Copper 20 Lead 14 Gallium 15

Using the discriminant function developed from the Sturgeon Lake data set, what is the probability that it could be assigned to reef carbonate facies? The posterior probability that the new well has penetrated the talus of a nearby reef is 0.946 (assuming a prior probability of 0.5). Renewed efforts should therefore be made to locate the reef and a new well drilled or the current hole deviated.

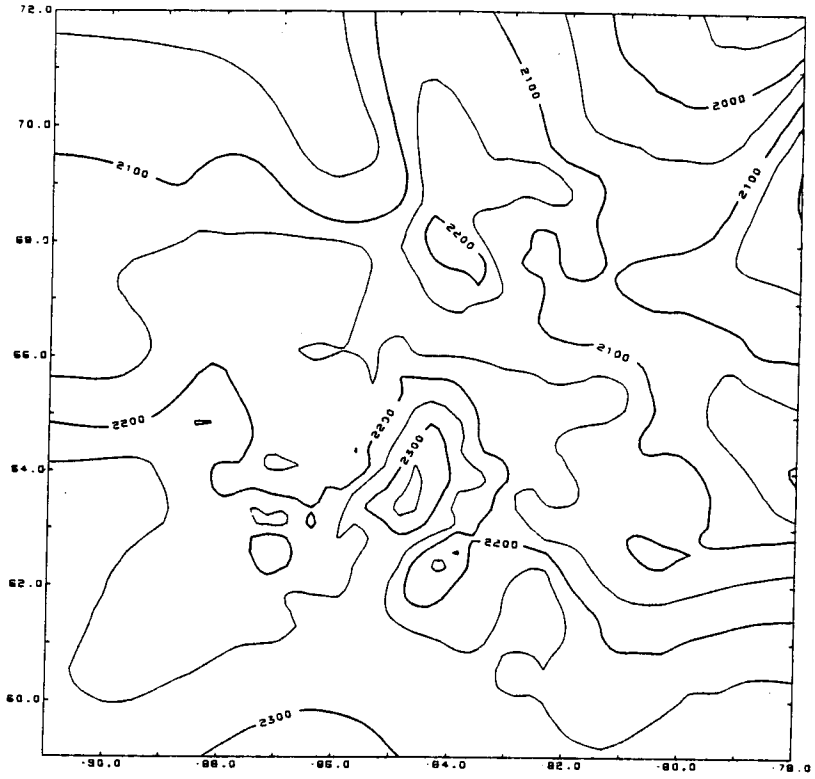
DISCRIMINANT FUNCTION ANALYSIS OF DRY HOLES AND PRODUCING WELLS: AN EXPLORATION STUDY IN THE MISSISSIPPIAN OF STAFFORD COUNTY, KANSAS

Small oil and gas fields are found in the Mississippian 'B' chert interval in South-West Stafford C unty, Kansas. Conventional exploration for these targets relies mainly on the location of structural highs by reflection seismic methods, although it is recognized that chances of finding a new field are enhanced in locations where the interval is relatively thick and has a low shale content. Data for three variables were assembled for a sample of 124 wells in the area which measure the following characteristics:

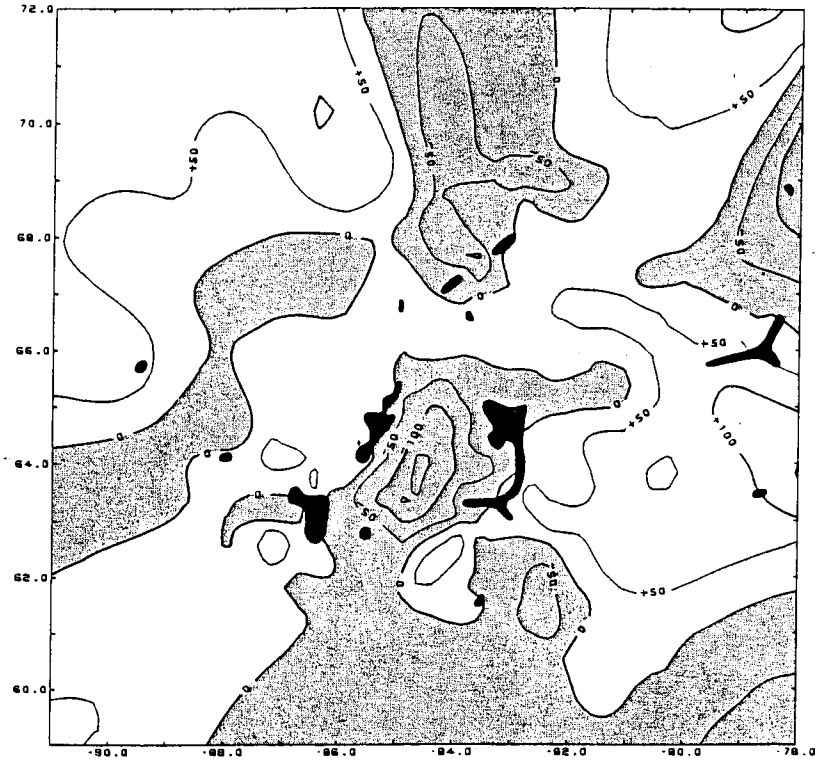
(1) The average gamma ray log deflection in the interval, standardized as a proportional measure of shale content, ranging between the extremes of zero (no shale) and one (total shale).

(2) The thickness of the interval (feet).

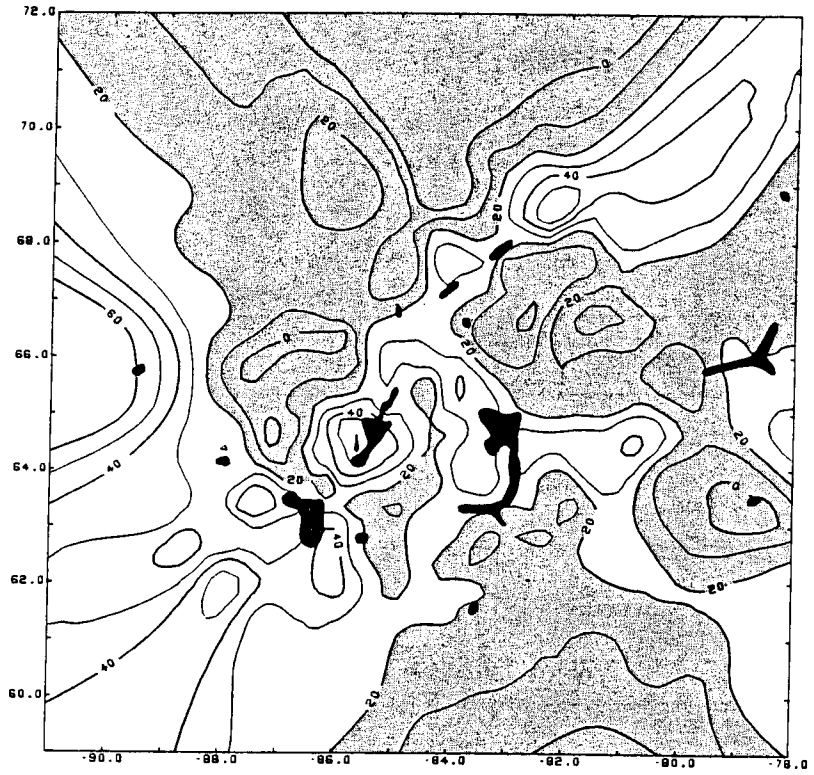
(3) Deviations from a linear trend surface (q.v.) of the structural elevation of the top of the interval as a measure of the local structure with simple regional dip removed (positive and negative elevations in feet).



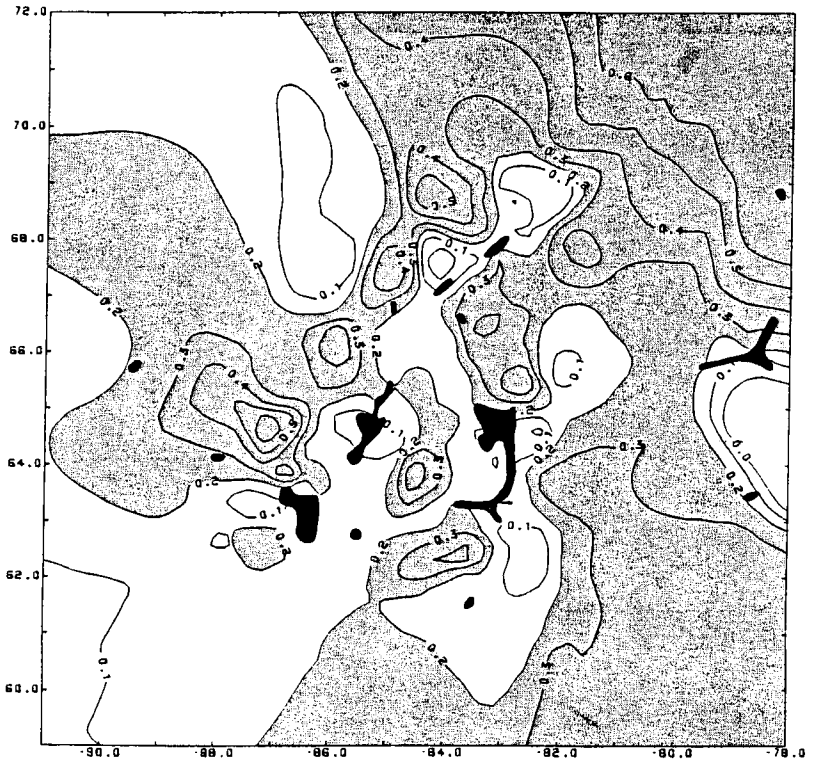
.--Structure of the top of the Mississippian 'B' (feet subsea).



.--Linear trend surface residuals of the Mississippian 'B' structure (feet subsea). Negative residuals are shaded; positive residuals are unshaded.

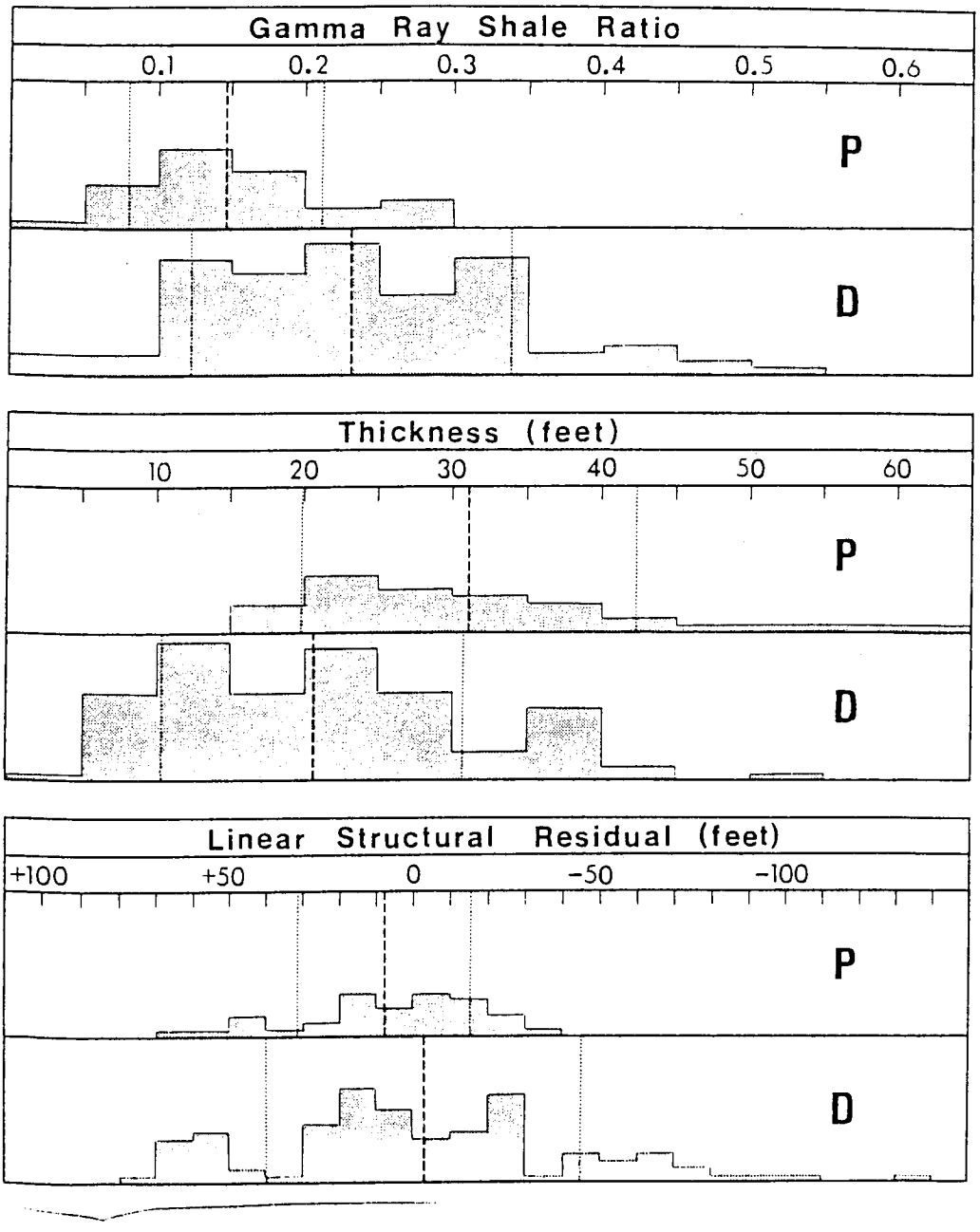


--Areal isopach development of the Mississippian 'B' zone (feet). Areas thinner than 20 feet are shaded; areas thicker than 20 feet are unshaded.



--Areal variation of the Mississippian 'B' gamma ray shale ratio. Areas with ratio greater than 0.2, shaded; areas with ratio less than 0.2, unshaded.

The sample was subdivided into a producing group P (33 wells) and a dry group D (91 dry holes). Histograms of the two groups for each variable are shown in the accompanying figure and indicate the degree of distinction between the groups in terms of each variable considered separately.



Frequency histograms of Mississippian 'B' shale ratio, thickness and structural residuals for producing (P) and dry (D) wells. Mean values are indexed by a heavy dashed line; boundaries about the mean of one standard deviation are indicated by dotted lines.

Computation of a linear discriminant function led to the equation:

$$z = -6.34\ G + 0.07\ T - 0.01\ S$$

where G = shale ratio

T = thickness

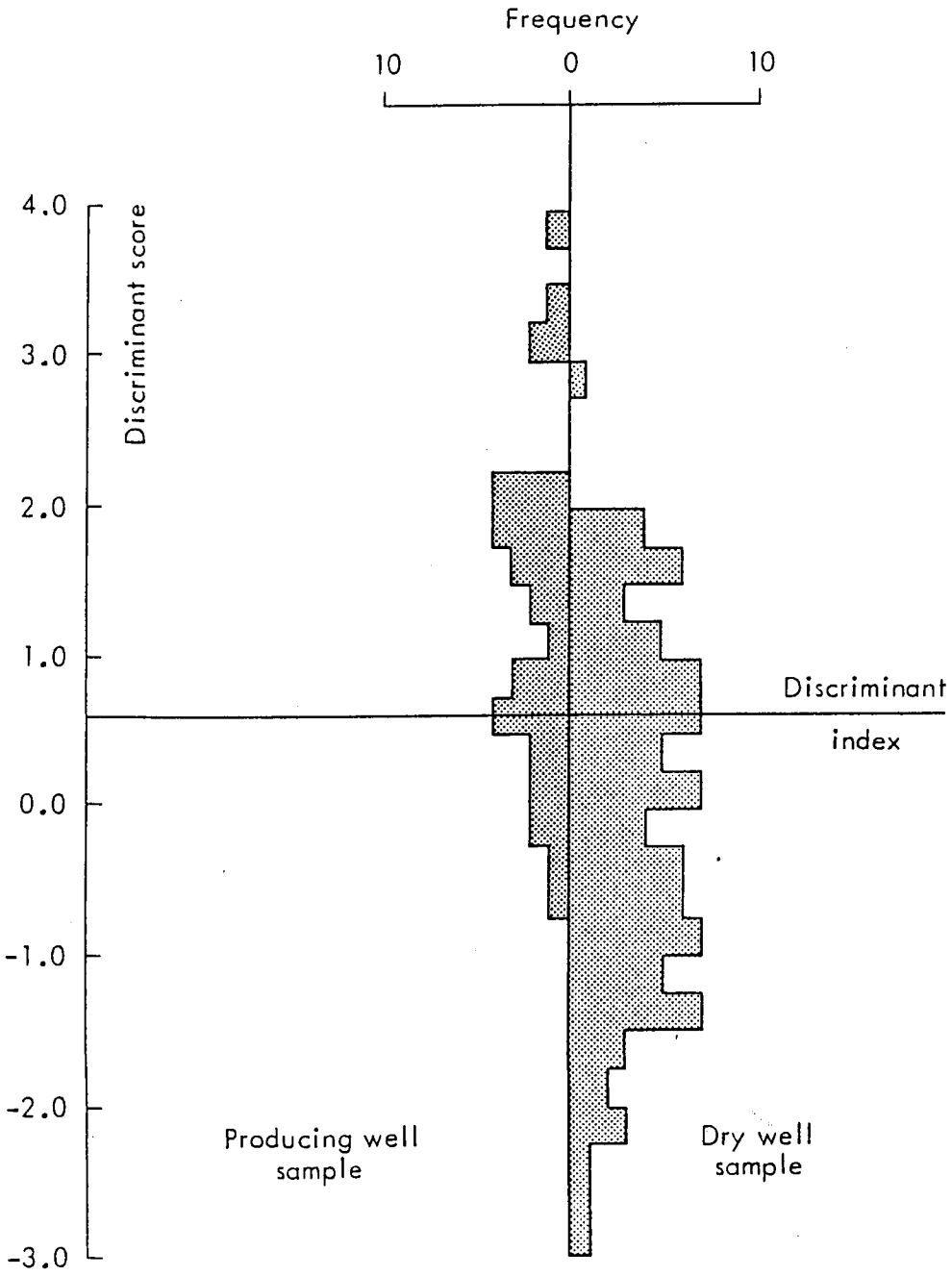
S = structural residual

The scores of the group centroids are:

$$\overline{z_P} = 1.30 \text{ and } \overline{z_D} = -0.08$$

where the discriminant index,

$$z_c = 0.61$$



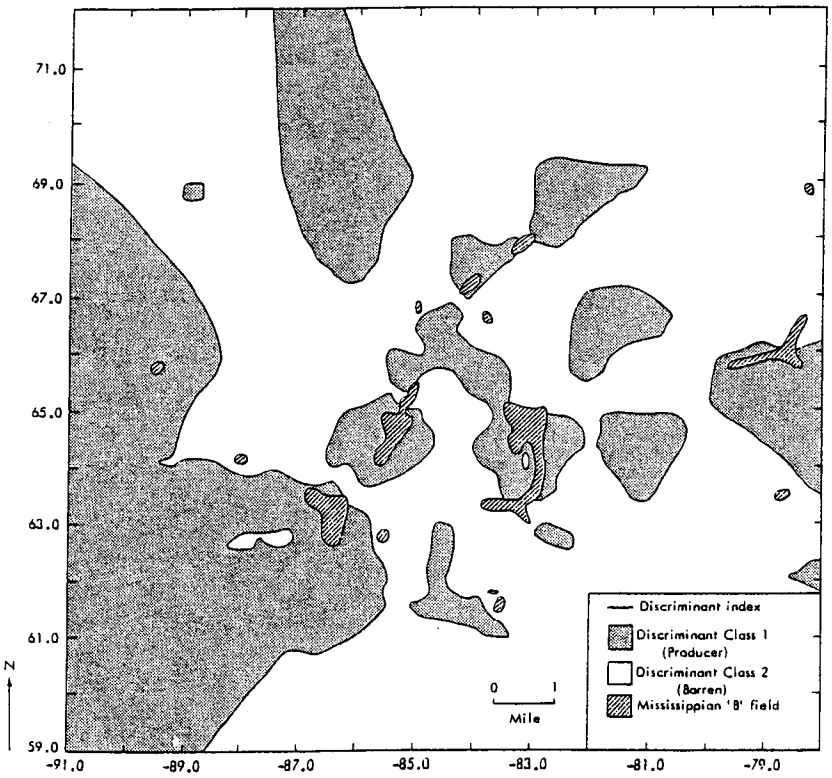
Contribution to discrimination by each variable as computed from their fractional component of Mahalanobis' distance are:

shale ratio, G:	39.6%
thickness, T:	52.6%
structural residual, S:	7.8%

The discriminant function analysis therefore suggests that the differentiation between dry and producing locales is made primarily in terms of thickness and shale content with relatively minor contribution by local structure. Producing wells are to be found in thick, "clean" chert sections and trapping contexts are stratigraphically, rather than structurally, controlled. An empirical measure of the success of the discriminant index as a critical classification boundary value between dry and producing wells may be made by tabulating the frequencies of correct classifications and misclassification.

DISCRIMINANT INDEX CLASSIFICATION

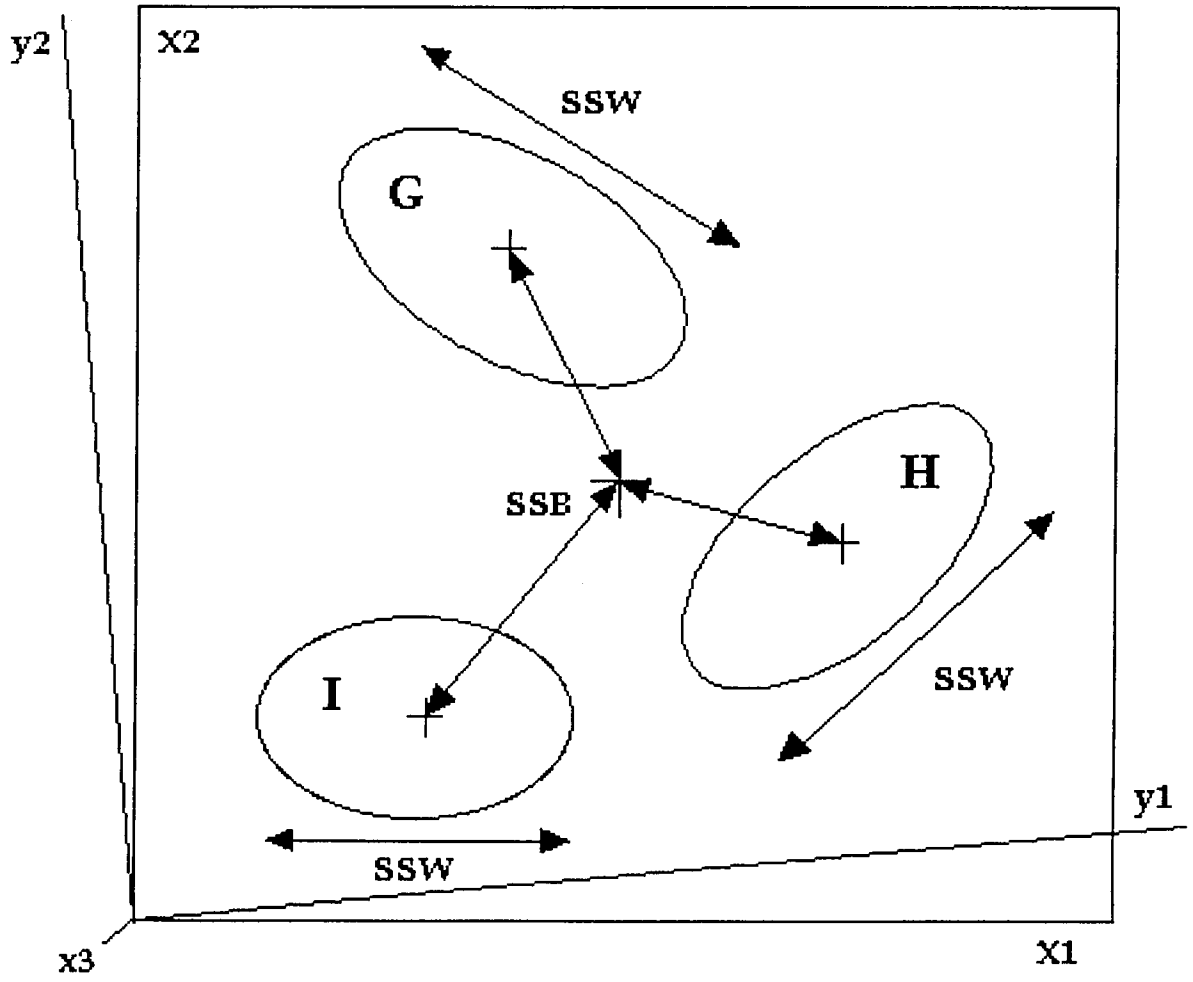
		Producer	Dry
Actual status	PRODUCER	23	10
	DRY	29	62



--Classificatory map of the study area in terms of producing and dry areas, based on the interpolation of computed discriminant scores.

Multiple Discriminant Analysis (MDA)

Multiple discriminant analysis (or canonical discriminant analysis) is a multi-group extension of linear discriminant analysis for two groups Instead of computing a single discriminant function, a solution is made for multiple functions which are orthogonal to each other. The first function locates the axis where the distances between the group centroids is maximized and the discriminant scores about the group means are minimized. The second function is orthogonal to the first and represents the second most powerful discriminator axis. The remaining functions account for successively lower discriminations. When observations are classified with respect to one or other of the groups, the assignment can be made in terms of probability, using a Bayesian estimation from the discriminant scores in conjunction with a prior probability. The prior probability is conventionally chosen to be either an equal probability for all groups or probabilities that correspond to proportion of observations to the total sample that is supplied by each group.



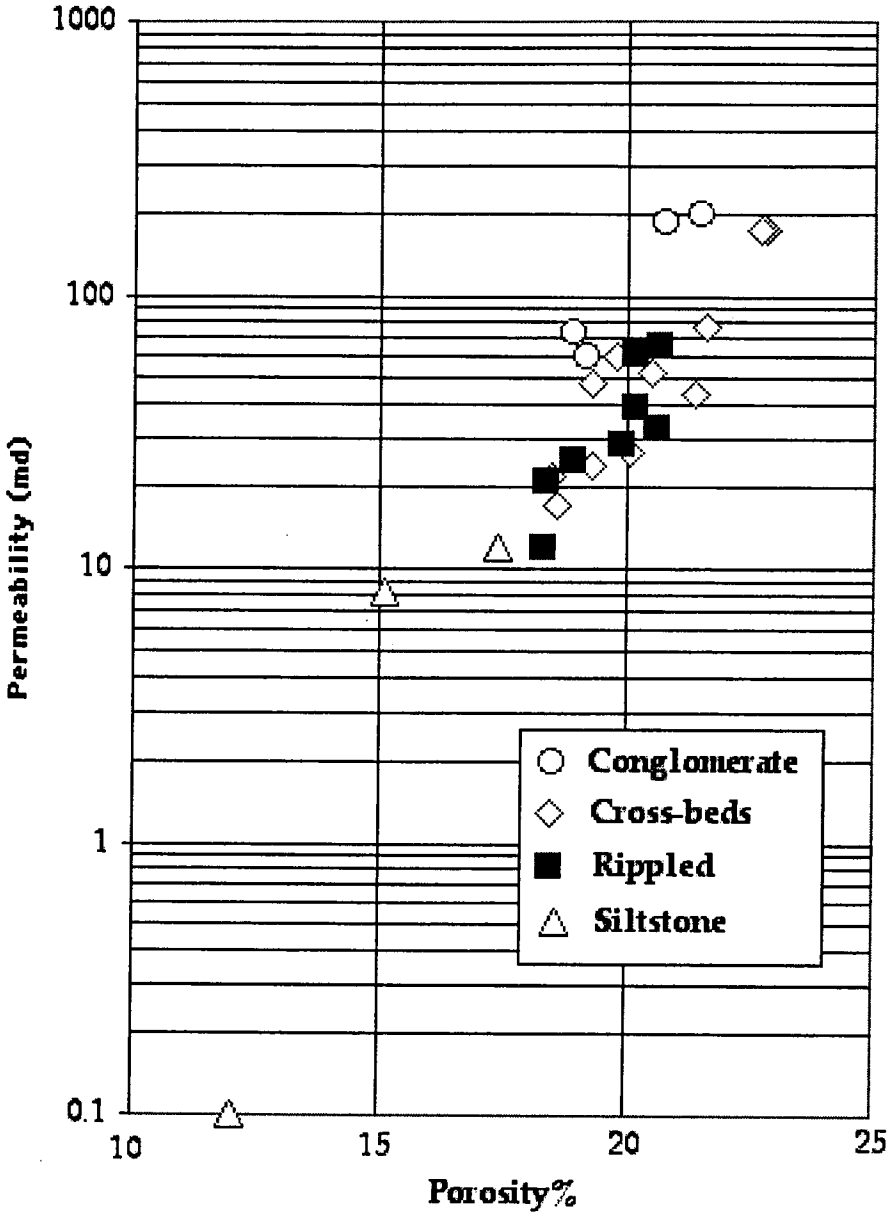
EXAMPLE: DISCRIMINATION BETWEEN CHEROKEE SANDSTONE DEPOSITIONAL TEXTURES SEEN IN CORE DETERMINED FROM POROSITY< PERMEABILITY, AND COMPOSITION

Elongate sandstone bodies of the Cherokee Group (Middle Pennsylvanian) in southeast Kansas have been prolific producers of oil for many years. Depositional origins of individual sands have been ascribed to nearshore bars, barrier islands, tidal channels, and alluvial valley-fill. Monroe 1-T (SE 9-21S-22E) was drilled through a productive section of the Skinner Sandstone in the Selma field. The reservoir section was cored. Porosity and permeability measurements were made on 2-inch plugs; bedding structures and textures described from examination of whole core; compositional estimates were made from microscopic examination of thin-sections. The results are tabulated below.

Monroe 1-T SE 9-21S-22E Skinner Sandstone, Cherokee, (Pennsylvanian), Selma field, Kansas						
DEPTH	POROSITY%	PERM (md)	QUARTZ	LITH	CEMENT	TEXT
714	17.4	12	46.6	21	15	SILT
715	12	0.1	38	50	0	SILT
716	19.9	29	55.1	20	5	RIPL
717	20.6	33	54.4	17	8	RIPL
718	20.2	62	56.8	18	5	RIPL
719	21.4	44	49.6	17	12	XBED
722	18.3	12	46.7	25	10	RIPL
723	18.9	25	41.1	35	5	RIPL
724	19.3	24	52.7	23	5	XBED
725	22.8	173	54.2	16	7	XBED
726	20.2	39	49.8	24	6	RIPL
727	18.4	21	39.6	37	5	RIPL
728	20.7	66	45.3	25	9	RIPL
729	22.7	173	45.3	18	14	XBED
730	20.1	27	47.9	23	9	XBED
731	18.6	17	35.4	35	11	XBED
732	21.6	78	35.4	29	14	XBED
733	20.5	53	34.5	30	15	XBED
734	19.8	61	28.2	38	14	XBED
735	15.1	8.1	32.9	38	14	SILT
736	21.5	197	40.5	18	20	CONGL
737	18.9	74	41.1	24	16	CONGL
738	19.2	60	41.8	23	16	CONGL
739	20.8	186	33.2	23	23	CONGL
740	19.3	48	37.7	28	15	XBED
741	18.5	22	38.5	28	15	XBED

The bedding structures and textures can be subdivided broadly into four classes: conglomeratic facies (CONGL), cross-bedded sandstones (XBED), ripple-bedded sandstones (RIPPL), and siltstones (SILT). Using multiple discriminant analysis, we will see how the rock fabric is related to porosity, permeability, and contents of quartz, lithic fragments, and calcite cement as a way to integrate geological and engineering data and characterize flow units for oil production.

A crossplot of porosity and permeability shows the broad trend of increasing permeability with increasing porosity. The conglomeratic facies has the highest permeabilities and the siltstones have the lowest permeabilities which reflects the marked difference in pore-throat sizes between the two rock types. The cross-bedded and ripple-bedded facies have porosity and permeability data clouds that overlap, although there is a tendency for the cross-bedded facies to have slightly higher porosity/permeability, which might be expected for larger grain sizes and pore-throats.



When a multiple discriminant analysis is run with the five variables of porosity, logarithm of permeability, quartz, lithology fragments, and calcite cement, for the four groups, the group means (and so the centroids of the group clouds in five-dimensional space) are:

Group	Count	PHI%	logk	QUARTZ	LITH	CEMENT
CONGL	4	20.10	2.05	39.15	22.00	18.75
XBED	11	20.42	1.69	41.76	25.91	11.91
RIPPL	8	19.65	1.50	48.60	25.13	6.63
SILT	3	14.83	0.33	39.17	36.33	9.67
Global	26	19.49	1.53	43.17	26.27	11.08

The pooled covariance matrix for the groups (reflecting the orientation and inflation of the group clouds in five-dimensional space) is:

	PHI%	logk	QUARTZ	LITH	CEMENT
PHI%	2.24	0.58	4.36	-8.83	2.23
logk	0.58	0.20	0.48	-2.09	1.03
QUARTZ	4.36	0.48	53.23	-46.50	-11.09
LITH	-8.83	-2.09	-46.50	61.38	-6.06
CEMENT	2.23	1.03	-11.09	-6.06	14.92

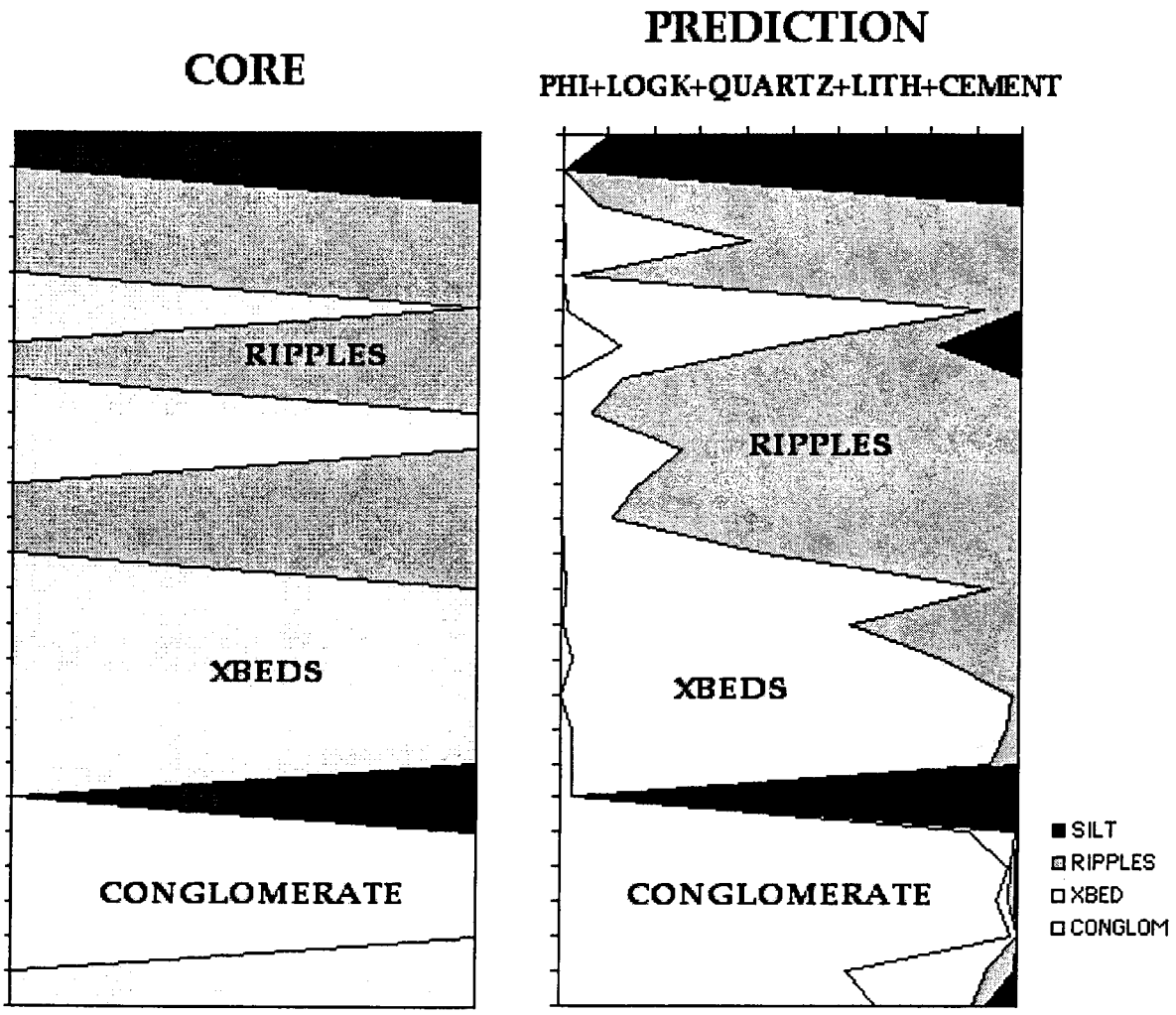
The multiple discriminant classification of the zones is then:

CONGL	XBED	RIPPL	SILT	PRED	CORE
0.096	0.001	0.000	0.903	4	4
0.000	0.000	0.000	1.000	4	4
0.001	0.078	0.921	0.000	3	3
0.005	0.402	0.593	0.000	3	3
0.001	0.018	0.980	0.000	3	3
0.008	0.910	0.082	0.000	2	2
0.127	0.347	0.339	0.187	2	3
0.000	0.133	0.867	0.000	3	3
0.001	0.064	0.934	0.000	3	2
0.000	0.263	0.737	0.000	3	2
0.001	0.171	0.828	0.000	3	3
0.001	0.111	0.888	0.000	3	3
0.004	0.436	0.559	0.000	3	3
0.008	0.927	0.065	0.000	2	2
0.006	0.622	0.371	0.000	2	2
0.024	0.809	0.166	0.001	2	2
0.002	0.981	0.017	0.000	2	2
0.024	0.952	0.024	0.000	2	2
0.025	0.911	0.064	0.000	2	2
0.026	0.000	0.000	0.974	4	4
0.891	0.105	0.004	0.000	1	1
0.979	0.007	0.009	0.005	1	1
0.952	0.027	0.014	0.007	1	1
0.984	0.016	0.000	0.000	1	1
0.625	0.306	0.067	0.003	1	2
0.687	0.213	0.031	0.069	1	2

The classification table is:

	PREDICTION			
CORE	CONGL	XBED	RIPPL	SILT
CONGL	4			
XBED	2	7	2	
RIPPL		1	7	
SILT				3

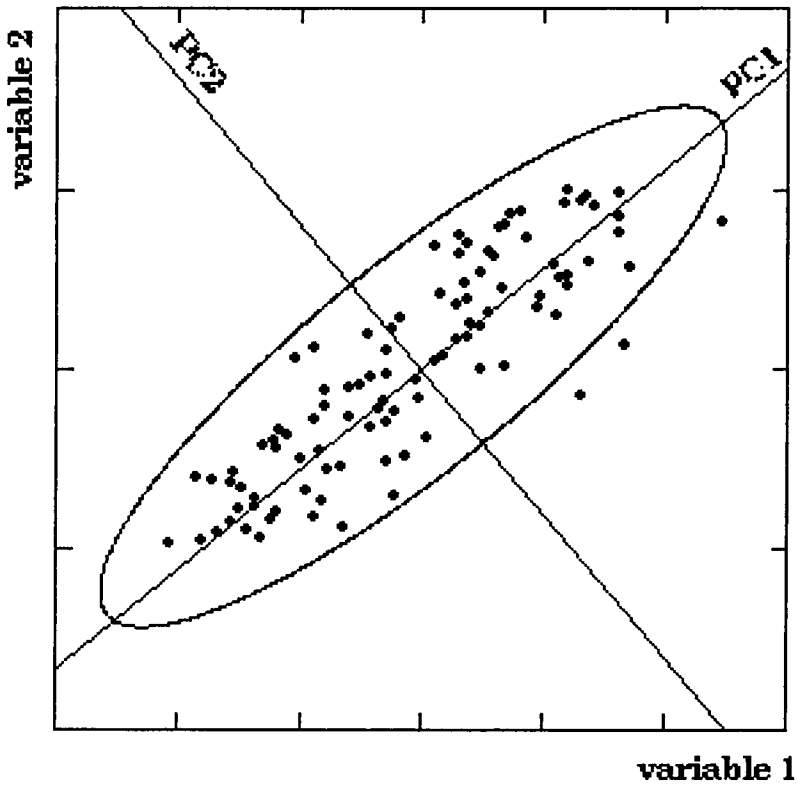
and a graphic comparison between the core classification and the discriminant prediction is:



Examination of the discriminant statistics reveals which variables are most strongly linked with discrimination between the bedding-type facies. The depth plot of probabilities also gives insights on vertical trends in changes in fabric properties as well as highlighting anomalous zones where the petrophysical / petrographic properties appear to be at variance with the expectation for their bedding type.

PRINCIPAL COMPONENT ANALYSIS (PCA)

If m variables for a sample of observations are plotted as points in a space with mutually orthogonal axes, they form a cloud in m-dimensional space. Principal components are the eigenvectors of this cloud, computed to locate the major axes in order of importance. These axes provide a new framework of reference which is aligned with the natural axes of the cloud ("eigen" is German for "intrinsic"), rather than the measurement variable axes. The orientation of the principal components are computed from either the covariance or correlation matrix of the data. The correlation matrix is the more common choice, because many observational variables are recorded in radically different units. In order to avoid artificial and undue weighting by any of the variables, the original data should be standardized to dimensionless units by subtracting the mean and dividing by the standard deviation. The covariance matrix of standardized data is the correlation matrix.



The raw data cloud is modelled by a single hyperellipsoid. The ellipsoid orientation is at some angle to the measurement axes and the relative directions of axis elongation reflect systematic relationships between the variables. The axes of the ellipsoid are the eigenvectors, whose relative magnitudes are given by their associated eigenvalues. A compression of dimensionality is made possible because the components are based on the intercorrelations between the

variables. As a simple and extreme example, if there is a perfect correlation between three variables, then data points will be strung out on a single oblique axis. So, although the points are found in a three-dimensional space of the original variables, the intrinsic dimensionality of the data variability is only one. In a more realistic case of high correlations between variables there will still be a strong tendency for data to be aligned along an axis, with the residual scatter of points absorbed by subsidiary and shorter axes. The first principal component therefore accounts for the maximum amount of variability of any single possible axis. The remaining principal components pick up the rest of the variability in an ordered allocation.

The total variance of the original set of m variables is the sum of their separate variances. This quantity is absorbed by m possible principal components. In practice, many measurement variables show a significant degree of intercorrelation, so that the last few principal components may account for trivial amounts of the total variability. Put another way, this property highlights the amount of information redundancy within the variables. If the majority of the variability is picked up by p principal components, then the dimensionality has been shrunk from m to p. The collapse reflects the dimensionality of the information content of the variables as a replacement for the original reference framework. It is not uncommon for the first two principal components to account for most of the variability, thus a multivariable data set can be mapped on a crossplot with little loss of information.

The derivation of the principal components follows from a property of matrix algebra that a symmetric, nonsingular matrix, $\mathbf{S}$, can be converted into a diagonal matrix, L, by multiplying by an orthonormal matrix, U, through the following equation:

$$\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U} = \mathbf{L}$$

where T signifies the transpose of a matrix. If $\mathbf{S}$ is the covariance matrix, then the conversion to a diagonal matrix is the geometrical equivalent of a rotation of the original axes to new descriptive axes. The diagonal matrix has zeroes in the off-diagonal elements, which means that the new axes are independent of one another. The values of the diagonal elements register the eigenvalues of these principal components which express their variances. The sum of these eigenvalues is then the same as the sum of the variances of the original variables. The relationship gives an immediate measure as to how much variability is assigned to each principal component. The numbers are particularly easy to follow when the correlation matrix is selected. The variance of each variable is then unity, and the total variability equals m (the number of variables). Each eigenvalue divided by m is the proportion of a principal component's share of the total variability.

The fact that $\mathbf{U}$ is an orthonormal matrix leads to the useful result that the inverse of $\mathbf{U}$ is the same as the transpose of $\mathbf{U}$. This means that both the transformation from the measurement space to principal component space and the reverse mapping are variations of the same operation. The matrix $\mathbf{U}$ contains the loadings that relate the eigenvectors to the original variables. The

location of any point within the data cloud can be related to the principal component axes by the transformation:

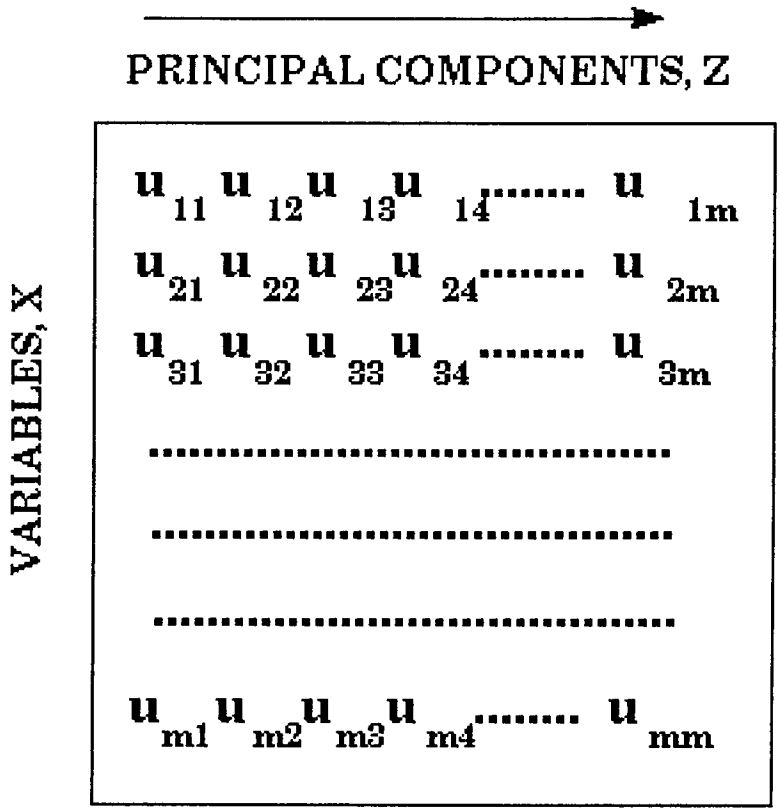
$\mathbf{Z} = \mathbf{U}^T\mathbf{X}$

where $\mathbf{X}$ is a vector of the variable responses and $\mathbf{Z}$ is a vector of principal component scores. This means that the score of the ith zone on the pth principal component is given more simply by:

$z_{pi} = u_{1p}x_{1i} + u_{2p}x_{2i} + \ldots + u_{mp}x_{mi}$

where the u coefficients are loadings from the pth principal component. These are the values of the pth <u>column</u> in the table matrix, shown diagrammatically in the figure. The original variables can be recovered from the principal component scores through the inverse of this procedure:

$\mathbf{X} = \mathbf{UZ}$



$\mathbf{X} = \mathbf{UZ}$; $\mathbf{Z} = \mathbf{U'X}$

So, the gth variable of the ith observation can be computed by the equation:

$x_{gi} = u_{g1}z_{1i} + u_{g2}z_{2i} + \ldots + u_{mp}z_{mi}$

which on the figure would correspond to the multiplication of the scores by the loadings of the gth <u>row</u>.

The ability to move easily between the two systems of coordinates is more than a mathematical convenience. Subsets of principal components can be mapped and related directly to the original variables. The loadings of the **U** matrix summarize the relationships between the variables and the principal components. As such, they can often be "read" for their meaning in terms of the original variables. The observational variables are indirect measurements of fundamental, but unseen, properties. In the PCA of the correlation structure of relationships, the principal components will often reveal these properties implicitly. Common sense must be used in such interpretations. The computation of principal components is simply a geometrical operation which relocates the reference axes to the apparent axes of elongation of the data cloud. The preceding explanation only covers the bare bones of the mathematics of principal component analysis. The ideas and further ramifications are best understood by consideration of actual examples.

The methods of discriminant analysis described previously are types of *supervised* pattern-recognition procedures. Training data sets are used to develop discrimination statistics that best partition the classes recognized before the analysis. The success of the discrimination is first checked by classification predictions from the training or calibration set. Whenever possible, a more realistic assessment is made by measuring classification performance in another known, but independent validation set.

By contrast, when principal component analysis is used to look for patterns and discriminations within data it operates in an *unsupervised* manner. The PCA solution is simply a geometric rotation in multidimensional space that locks onto the orthogonal axes of relative elongation in a cloud of data points. The cloud is represented by a covariance or correlation matrix, so that the derived principal components respond to the structure of intercorrelations between the measurement variables. Any correlations other than zero indicate that there is some degree of information redundancy in the system and that the full dimensionality can be collapsed to a representation in fewer dimensions. In many instances the majority of the variability is allocated to the first two components, in which case a crossplot will show the basic structure of a multidimensional cloud. This attribute of dimensional reduction is a major reason for the popularity of principal component analysis in data processing.

CASE STUDY: ELECTROFACIES ANALYSIS OF A MISSISSIPPIAN COMPLEX CARBONATE UNIT

In this case-study, a Mississippian carbonate section was logged but not cored, so that information on lithofacies is restricted to generalized drill-cuttings observations. In addition, the section is composed of variable amounts of chert, calcite, and dolomite. The neutron, density, and photoelectric factor logs from this well were used in the Compositional Analysis section of this manual to evaluate mineral and porosity content both from visual interpretation of the logs and as a matrix algebra solution using digital log data.

In an alternative strategy, electrofacies can be extracted from the unit by a pattern recognition approach to the logs, with interpretation that follows the analysis, rather than precedes it. Principal component analysis of the gamma ray, neutron porosity, density, and photoelectric factor logs are summarized in the table. The data show that the total variability in four-log space can be mapped almost entirely in the two dimensions of the first two principal components which account for 97% of the total variance. The loadings of the first principal component show the strong distinction between the high-porosity cherts and lower-porosity carbonates. The second principal component appears to be keyed principally with gamma-radioactivity.

While no core information is available to locate reference lithofacies in these log clouds, electrofacies can be differentiated by pattern recognition methods. A quick method is to locate three endmember points as vertices of a compositional triangle and then solve the composition of any zone as proportions of these three endmembers in a methodology which is a simple adaptation of the compositional analysis method discussed earlier.

PRINCIPAL COMPONENTS ANALYSIS
OF CHAT SECTION

VARIANCE EXPLAINED BY COMPONENTS

1	2	3	4
3.101	0.778	0.078	0.043

PERCENT OF TOTAL VARIANCE EXPLAINED

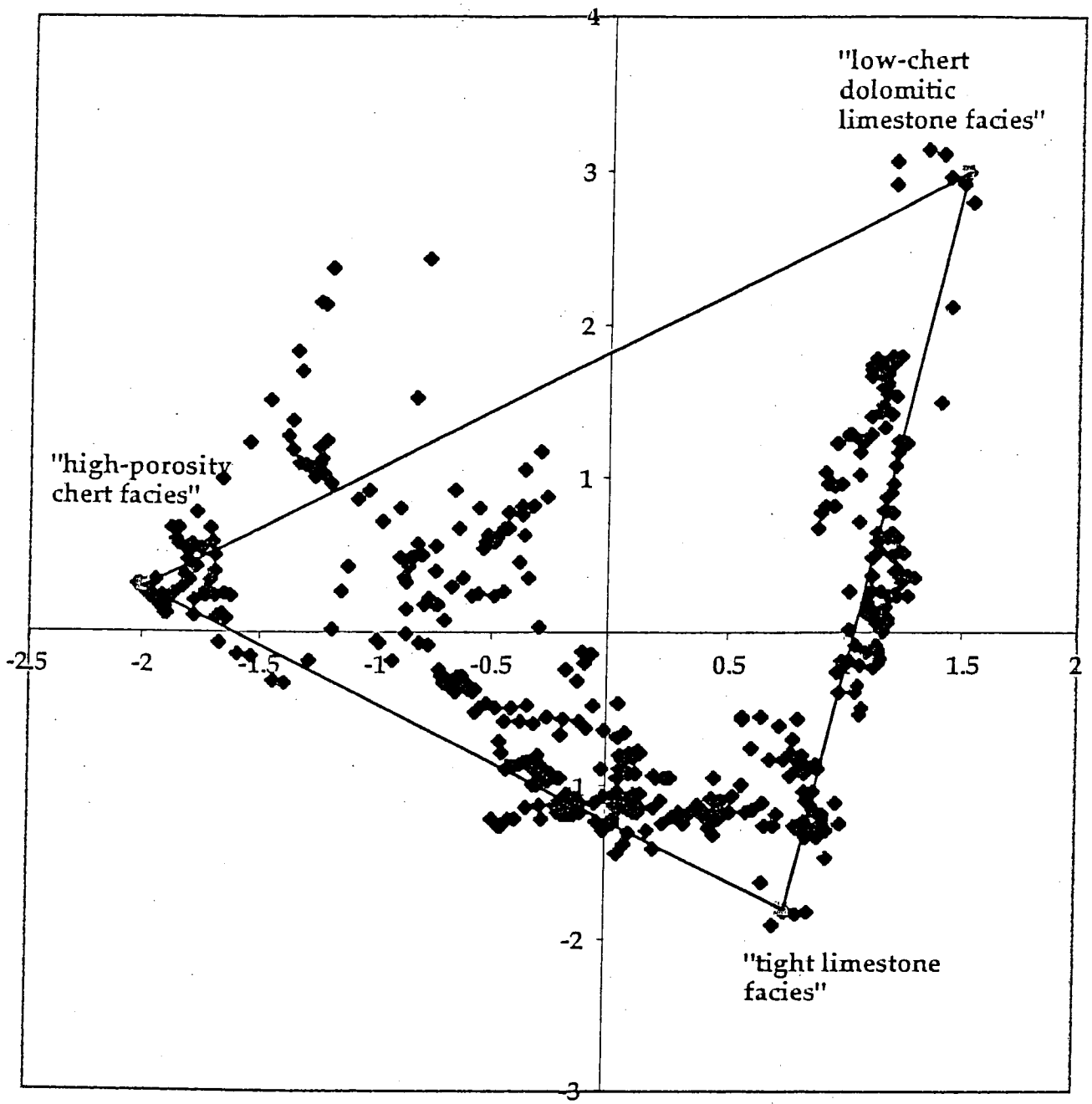
1	2	3	4
77.516	19.457	1.962	1.065

COMPONENT LOADINGS

	1	2	3	4
GR	0.573	0.818	-0.043	-0.021
PEF	0.971	-0.048	0.233	0.001
DPHI	-0.97	0.158	0.107	-0.149
NPHI	-0.942	0.285	0.104	0.142

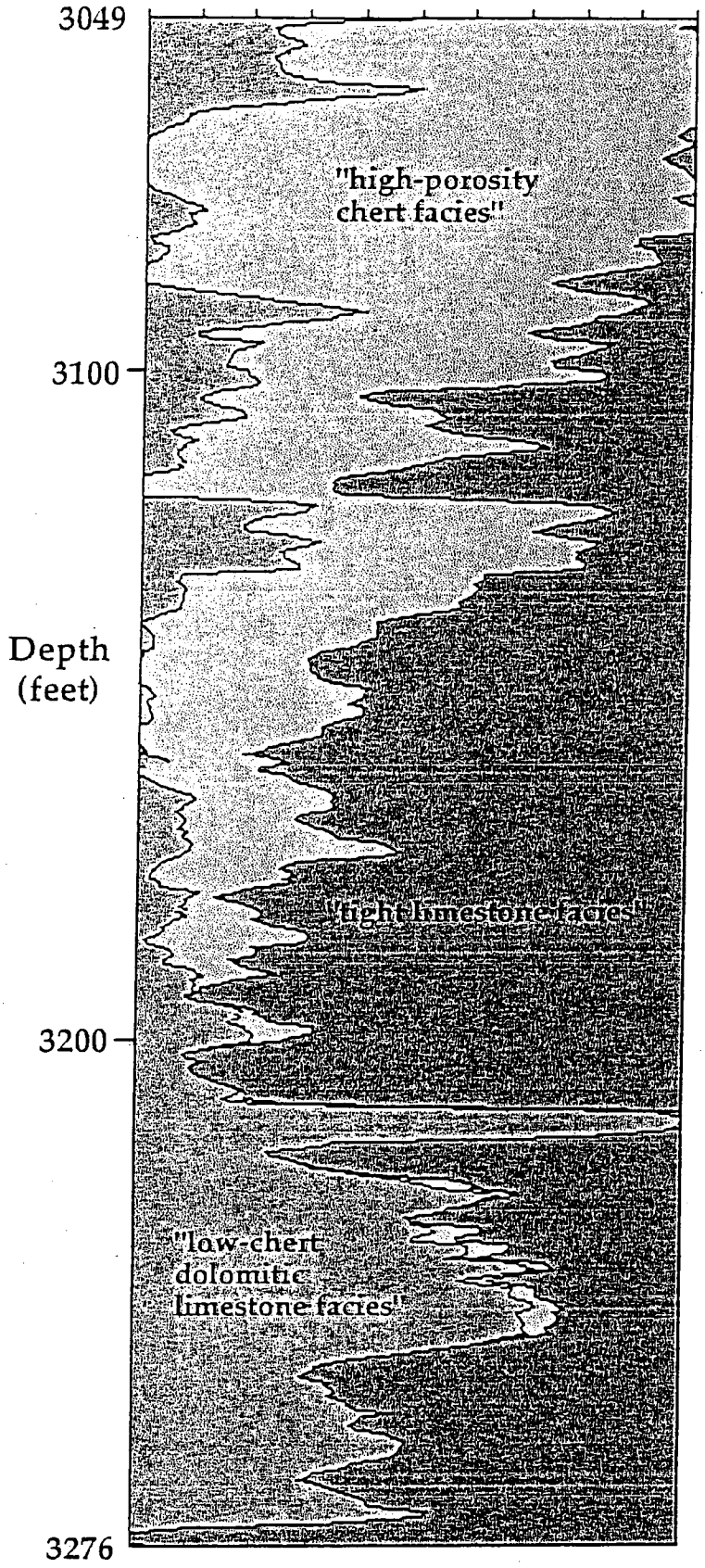
COMPONENT SCORE COEFFICIENTS

	1	2	3	4
GR	0.185	1.051	-0.55	-0.484
PEF	0.313	-0.062	2.971	0.023
DPHI	-0.313	0.203	1.361	-3.49
NPHI	-0.304	0.366	1.327	3.324



Pioneer Exploration Petrie #4 SE-NW-SW 36-26S-1W Sedgwick Co., Kansas

Crossplot of the first two principal components of the gamma ray, photoelectric factor, neutron and density porosities referenced with vertices of a composition triangle that appear to locate three electrofacies endmembers.



Pioneer Exploration Petrie #4 SE-NW-SW 36-26S-1W Sedgwick Co., Kansas

Electrofacies profile plot of Mississippian carbonate section computed as compositions of reference facies located on the crossplot of the first two principal components of the gamma ray, photoelectric factor, neutron and density porosities.

THE POISSON DISTRIBUTION

The Poisson distribution describes the expected frequencies of events per unit area or length, provided that these events are scattered randomly and are comparatively infrequent in occurrence. A distinct advantage of the Poisson distribution is that it may be applied in situations where the quantity *np* is known, but neither n (the number of trials) or p are known independently. This condition occurs in a wide variety of applications where "rare" events are observed to occur within time periods, areas and volumes of fixed size and their non-occurrence ("failure") is not ennumerable.

The Poisson distribution is defined by the Poisson probability P(X):

$$P(X) = \frac{e^{-np}(np)^X}{X!}$$

where e = 2.718 and X = the number of events in an interval.
The estimate of the mean is

$$\overline{X} = np$$

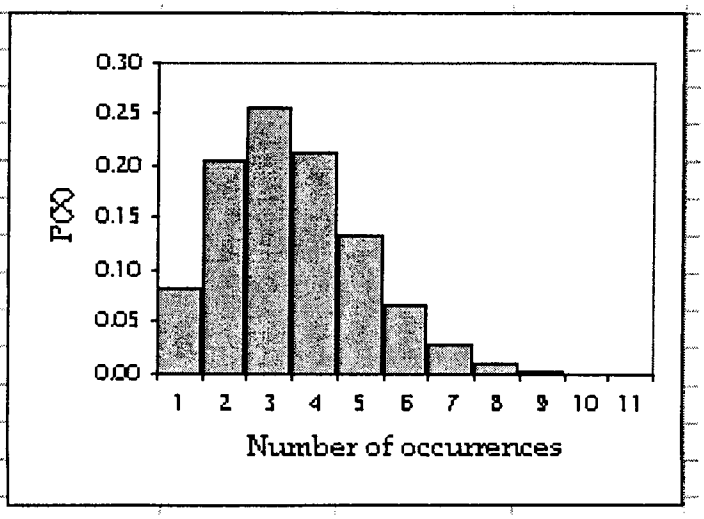
which is the mean number of events per interval or the average rate of occurrence.
The value of mean is the same as that of the variance:

$$s^2 = np$$

The Poisson distribution is the limiting approximation of the binomial distribution for low values of p, where the binomial distribution becomes highly skewed. Notice that the mean value for the Poisson distribution is the same as for the binomial and that the Poisson variance is the limiting case of the binomial s^2=npq as q approaches unit value. The Poisson distribution can be generated easily in EXCEL using the function POISSON(X, mean, 0) where X is the number of events, mean is the mean value (np), and the zero is a FALSE designation, so that the probabilities for each occurrence frequency is generated (as distinct from a one for TRUE, which would give a cumulative probability). So, for example, when the mean value is 2.5 number of occurrences per unit, then the Poisson distribution is:

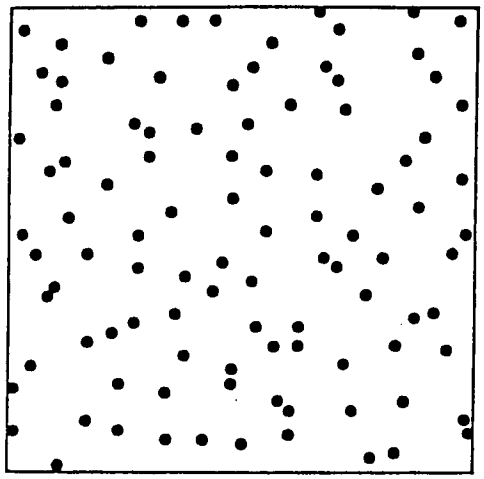
np=2.5

X	P(X)
0	0.0821
1	0.2052
2	0.2565
3	0.2138
4	0.1336
5	0.0668
6	0.0278
7	0.0099
8	0.0031
9	0.0009
10	0.0002

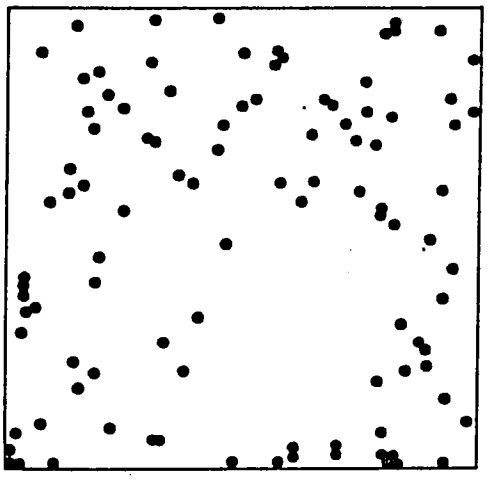


Notice that the Poisson distribution was generated using only one parameter, the mean, because the variance takes the same value as the mean.

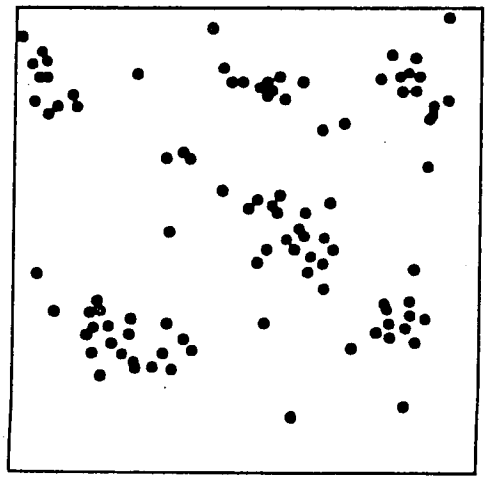
We can count the number of occurrences per unit, compute the mean n umber of occurrences and their variances. If the occurrences occur randomly, then we expect the distribution to be matched by the Poisson and the mean to be approximately the same as the variance. If the variance is markedly lower than the mean, then the distribution is more uniformly distributed than we would expect to see for a random process; if the variance is distinctly higher than the mean, then the distribution is more clustered.



Uniform distribution



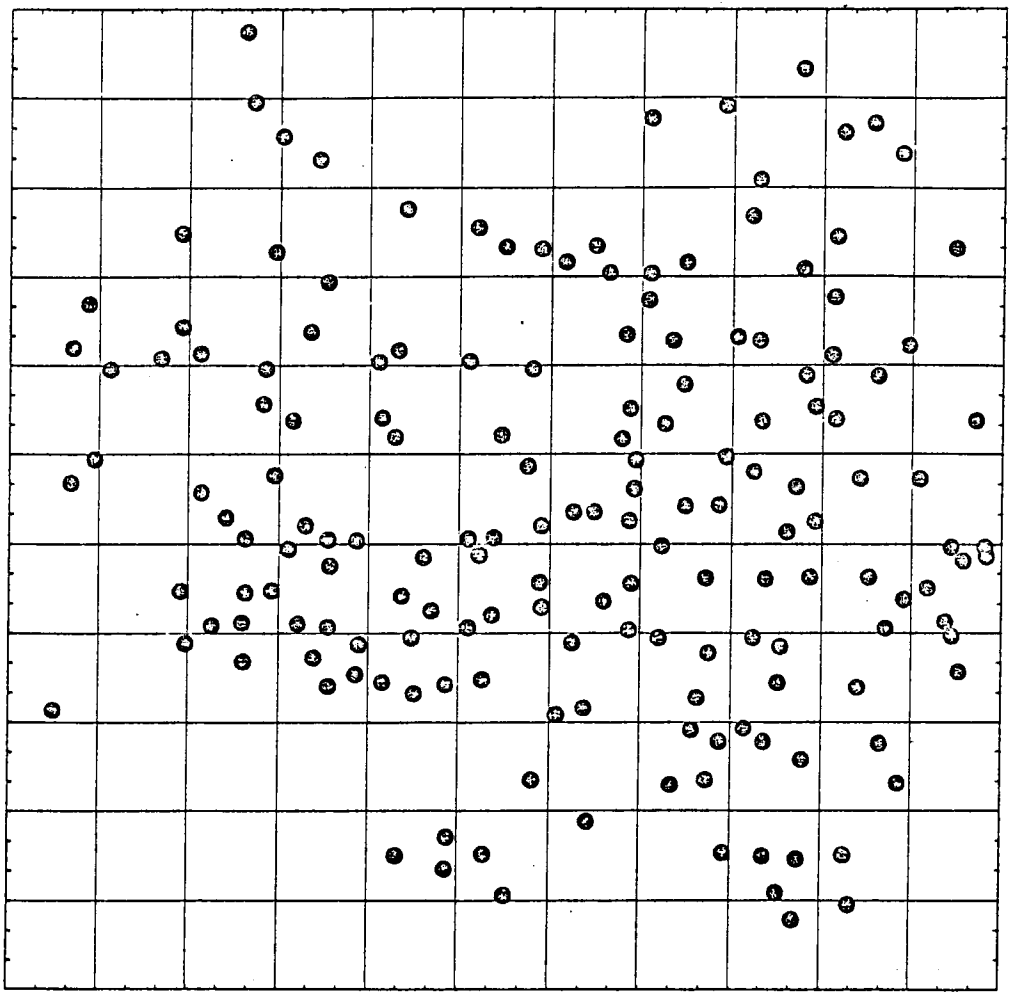
Poisson (random) distribution



Contagious (clustered) distribution

LANSING DISCOVERY WELLS IN NORTHWEST KANSAS

As an example of determining whether a spatial distribution of events is random, more uniform than random, or clustered, let us analyze the distribution of discovery wells in the Lansing Formation of a 33x33 mile area in northwest Kansas, as shown below.

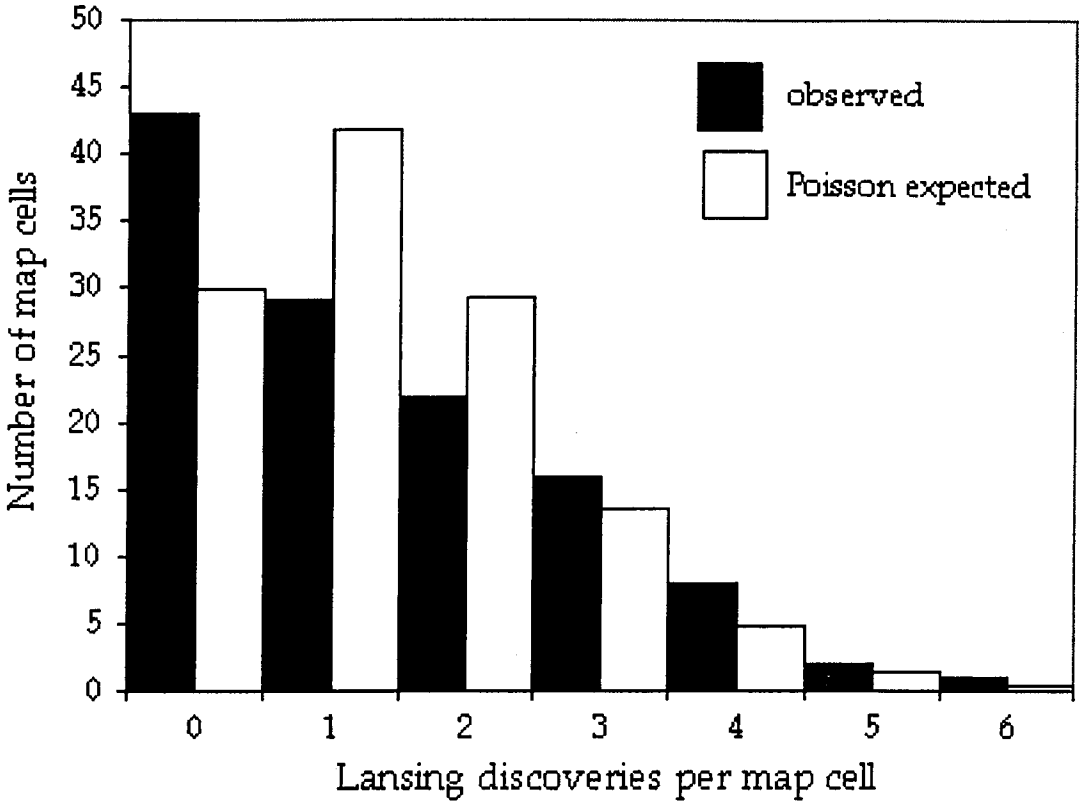


LANSING DISCOVERY WELLS

The area has been overlaid with an 11x11 grid which defines 121 map cells. There are 169 discovery wells, so the mean number of discoveries is 1.4 per map cell. The variance of the number of discoveries per map cell is then:

$$s^2 = \frac{\sum_{i=1}^{m} (X - \bar{X})^2}{(m - 1)}$$ where m is the number of map cells (121). The variance is calculated to be 1.97, which is greater than the mean value, so the initial indication is that the discoveries are clustered rather than random.

Using the EXCEL POISSON function, a histogram of freqencies expected from the Poisson distribution can be compared with the distribution of observed frequencies:



Comparison of the distributions show there to be a higher number of map cells with either zero or greater than 2 discoveries in the northwest Kansas area than would be expected if they were distibuted at random. This observation confirms the clustering phenomenom suggested by the variance being greater than the mean. With the expected value for the mean:variance ratio to be unit value for the Poisson distribution, how much higher or lower than unity must the ratio be in order for us to reject the null hypothesis that the distribution is random? This question can be resolved by a t-test.

The mean:variance ratio for the discovery wells is: 0.70. This is a sample estmate of the ratio, based on 121 observations. The null hypothesis postulates that this estimate is distributed about the Poisson parameter mean:variance ratio as a t-distribution, where the spread is dictated by the standard error, s_e .

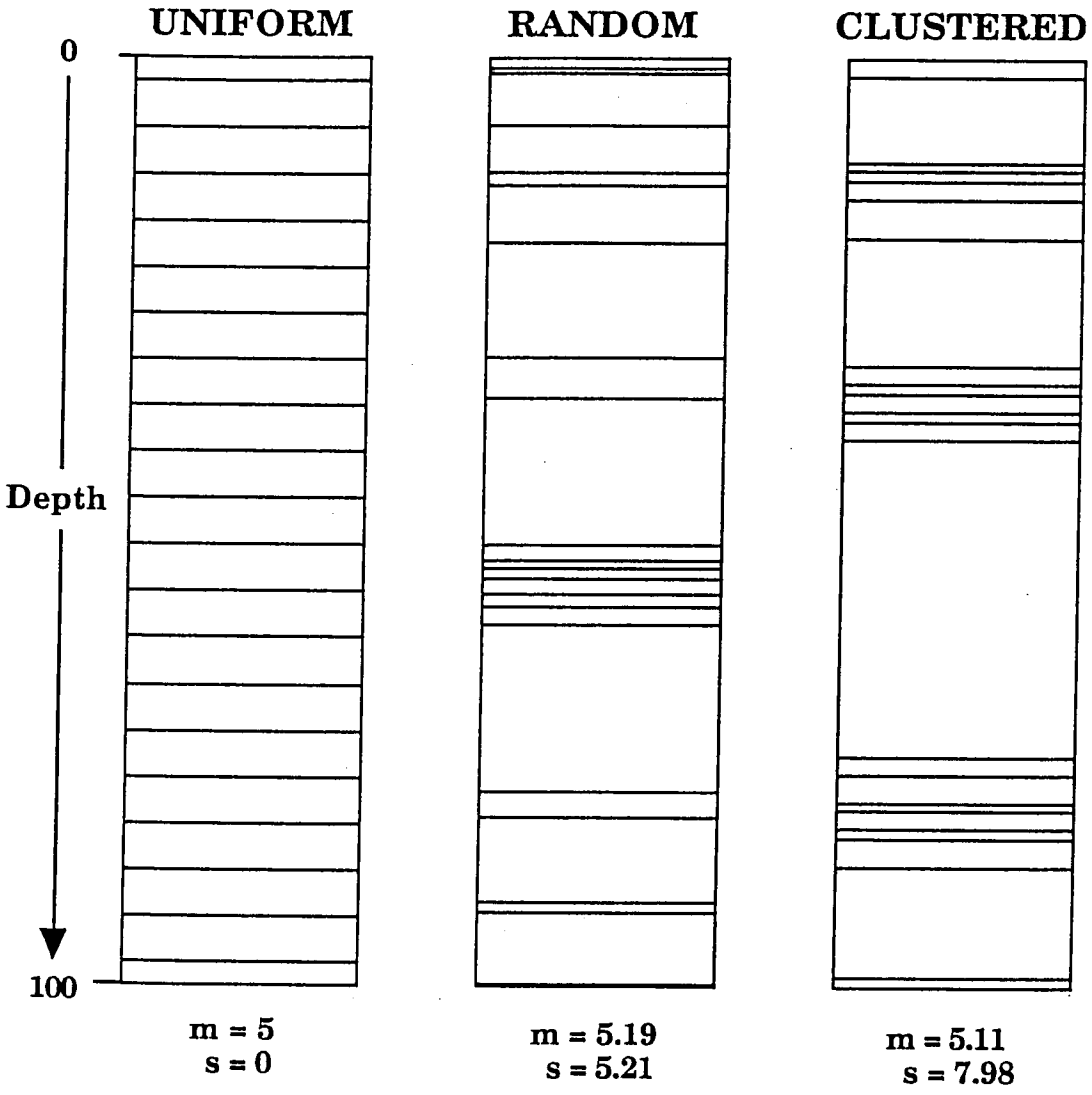
The standard error for the Poisson distribution is given by: $s_e = \sqrt{\frac{2}{m}}$

which is calculated to be 0.129.

The t-statistic is then $t = \frac{(1 - ratio)}{s_e}$ which gives a value of 2.32 This value exceed the critical t-test value at 120 degrees of freedom (m-1) and a significance level of 0.05 which is 1.98. The null hypothesis of a random distribution is therefore rejected and the alternative hypothesis of clustering of discovery wells is accepted.

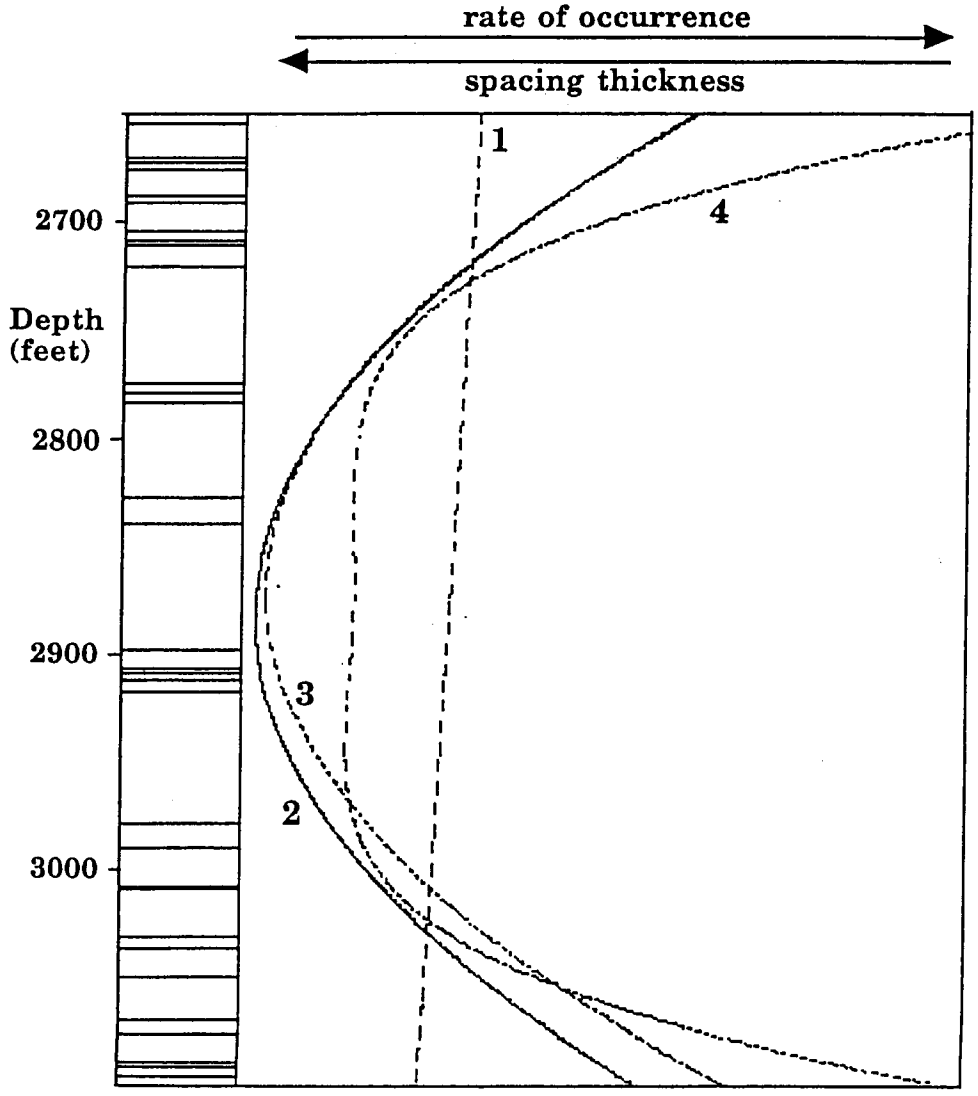
POISSON DESCRIPTOR MODELS FOR EVENTS IN TIME (OR ALONG AN AXIS)

As with the distribution of events in space, it is often difficult visually to distinguish random from clustered events on a time record. The human eye (or more accurately, brain) tends to see patterns of clusters and sparse occurrences in the results of random processes. This can be seen below, where the random occurrences in time are easily rationalized in terms of pattern.



A CASE STUDY: CHASE GROUP ANHYDRITE OCCURRENCES

Beds of anhydrite occur commonly throughout the Chase Group in southwestern Kansas and mark evaporite deposits thought to be linked with sabkha-like environments. Their relative depth positions may be useful zones marking exposure events in the depositional history of the group. The locations of anhydrite beds can be found through inspection of neutron–density log overlays or from *RHOmaa–Umaa* crossplots. A record of anhydrite occurrences in a Chase Group sequence is shown as the strip log below. A total of 34 anhydrite zones occur over a range of 450 ft. When the succession was subdivided into 18 equal intervals, it was found that the mean number of anhydrite zones per interval was 1.89 and the variance was 2.81. Based on the preceding discussion, it appears that the distribution is more clustered than would be expected if they were randomly distributed. However, some allowance must be made for the relatively small sample of anhydrite beds.



The predictions from a Poisson model of random events presume that the mean rate of occurrence is a constant. Put in another way, the depth record is presumed to be "stationary." If there is a systematic drift in the mean rate, then this is a trend that underlies the event sequence and it may have long-term geological significance. The nonstationary alternative can be modelled as a Poisson process with trend. Rather than work with frequencies of events in each (arbitrary) interval, it is often more convenient to consider the distance between successive events. Then the average distance, h, is given by the reciprocal of the rate of occurrence, r, that is:

$$h = \frac{1}{r}$$

Now, if there is no trend, the rate r is a constant, and is conventionally written as:

$$r = e^{a}$$

where a is a constant. If there is a linear trend with time, then:

$$r = e^{a_0 + a_1 t}$$

where t represents time, or, as in this example, depth. Taking logarithms of both sides:

$$\log r = a_0 + a_1 t$$

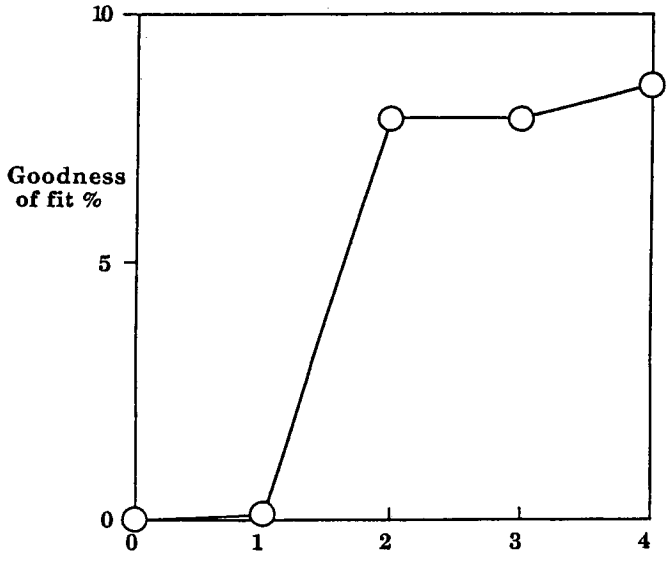
from which it follows:

$$-\log h = a_0 + a_1 t$$

because the length between events is the reciprocal of the occurrence rate. In fact, the trend can be expanded to a polynomial series in a very similar fashion to that used for trend analysis of continuous log data earlier in this chapter. For the mth polynomial the trend equation is:

$$-\log h = a_0 + a_1 t + \ldots + a_m t^m$$

The polynomial equations are the basis for regression analysis of the thickness between each pair of successive events versus the depth of the midpoint between the events. Consequently, the analysis does not require any special programming, but can be run using a conventional multiple regression procedure.



The results of fitting first- through fourth-order trend polynomials to the Chase Group anhydrite data are shown by the curves. A quadratic trend appears to be the best overall descriptor. This conclusion is based both on the curve shapes and the abrupt break in fit that is graphed . The degrees of fit are comparatively low and fail to pass muster as significant trends, as shown by the analysis of variance. This reflects the fact that the sample size is fairly small (33 occurrences of anhydrite), so that the high degree of fluctuation in successive spacings between anhydrite beds may overwhelm any systematic trend. However, notice that the variance ratios of the ANOVA table also indicate that the quadratic trend is distinctive. Strictly speaking, the aggregate spacing of events taken at a minimum of four at a time is preferable in order to average out some of the extremes (Cox and Lewis, 1966). However, this step becomes impractical for short sets of events and may mask some changes in occurrence rate.

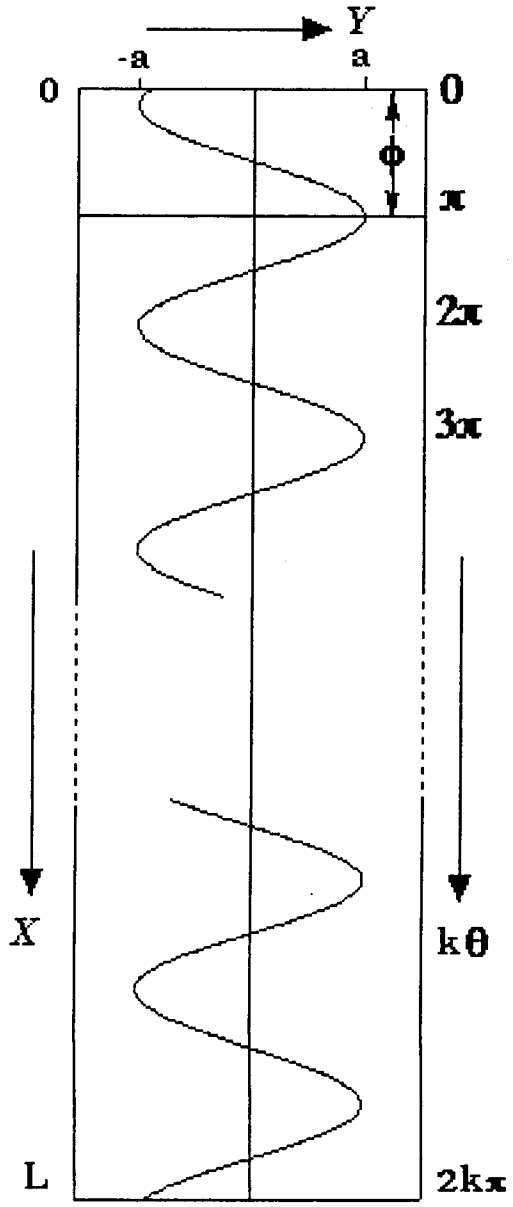
Source of Variation	Sum of Squares	DF	Mean Squares	F-ratio
Linear regression	0.5	1	0.5	0.1
Linear deviation	333.0	32	10.41	
Quadratic–Linear	26.0	1	26.0	2.9
Quadratic deviation	307.1	31	9.9	
Cubic–Quadratic	0.0	1	0.0	0.0
Cubic deviation	307.1	30	10.2	
Quartic–Cubic	2.5	1	2.5	0.2
Quartic deviation	304.6	29	10.5	
Total variation	333.5	33		

Critical F-test value at 5% significance, 1 and 31 df = 4.16

Because the functions used to describe the trends are polynomials, the depth locations of maxima, minima, and inflection points can be located using the methods described earlier. These depths provide global estimates of the overall distribution of events that can be interpolated between wells in the mapping of lateral changes in event rates.

ANALYSIS OF CYCLIC PATTERNS

A common objective in the interpretation of data recorded in time is the location of distinctive repetitive sequences or "cycles". In geological or petroleum engineering studies, wireline logs are particularly well-suited for the analysis of possible cycles, because such logs are numerical, lengthy, and continuous. Generally they contrast with records from outcrop and core successions, which are generally descriptive, short, and discrete. The fundamental model for cyclicity is that of the sine wave which sketches out the operation of a circular process as it develops through time. A repeating sine wave may be fitted to log data by a least-squares method that is a minor modification of conventional linear regression. This is shown in the figure, and is developed using some simple trigonometry as discussed in the following pages.



Both sine and cosine functions generate sinusoidal waves. The cosine equation to generate a wave is:

$$Y = a \cos \theta$$

where a is the amplitude of the wave and θ is an angular measure that changes with time in a cyclical manner. The wave will repeat at an interval set by its frequency, which is the number of cycles per unit of time or distance.

Because we are working with logs, we will take depth measurements as some monotonic function of time. The conversion from depth to angular measure can be understood from examination of the figure. If k cycles occur over a total thickness of h, and each cycle has a wavelength of p, then:

$$h = kp$$

If the depth axis is symbolized as X, then the conversion from depth to angles is given by:

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

when expressed in radians, where 2π represents one complete revolution.

Because radian measures are cyclic, the value of Y will repeat at depths that are displaced by multiples of p depth units. The cosine equation as written above is a limiting case, because it equates a depth of zero with an angle of zero and a single cycle. In order to make the equation more general for k cycles that originate at some unknown depth, it can be modified to:

$$Y = a \cos \left(k\theta - \phi\right)$$

where ϕ is called the phase angle and represents the initial shift displacement between the depth origin and the beginning of the nearest wave.

Now, from simple trigonometry, this can be expanded:

$$Y = a \cos \phi \cos k\theta + a \sin \phi \sin k\theta$$

Because the phase angle is a constant, the equation can be consolidated:

$$Y = A \cos k\theta + B \sin k\theta$$

where the tangent of the phase angle is equivalent to:

$$\tan \phi = \frac{B}{A} = \frac{a \sin \phi}{a \cos \phi}$$

The power or variance of the sinusoidal wave is the square of its amplitude, which is given by the sum of the squares of the coefficients:

$$A^2 + B^2$$

The expanded equation:

$$\hat{Y} = A \cos k\theta + B \sin k\theta$$

can be used to fit k cycles to a log trace, Y, by least squares in a conventional multiple regression of the general form:

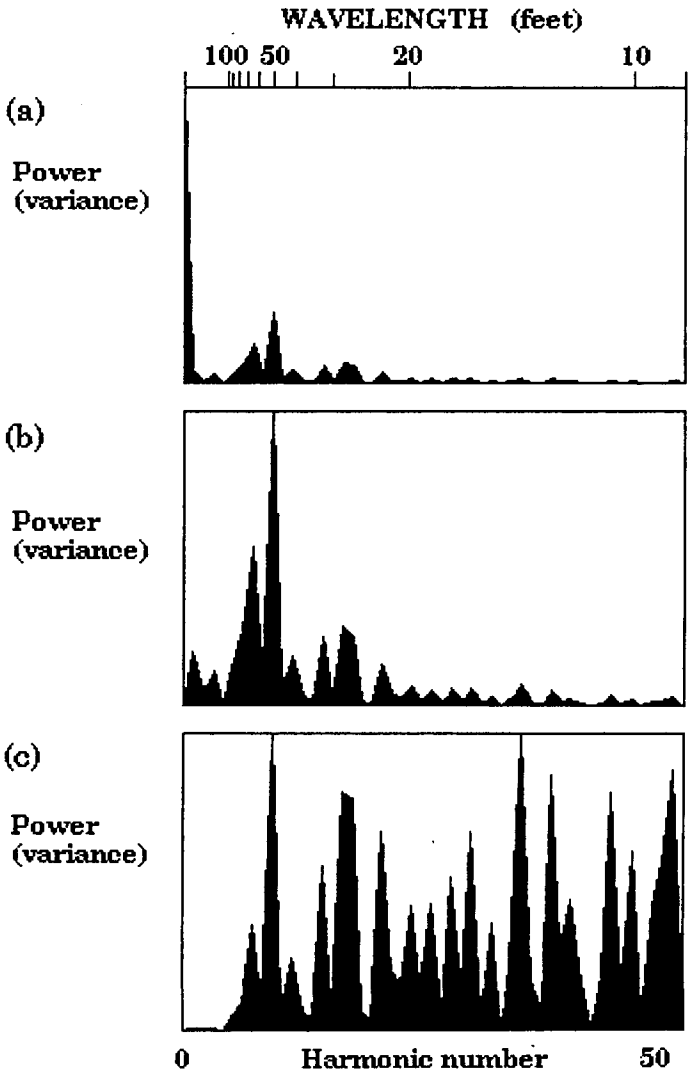
$$\hat{Y} = AX_1 + BX_2$$

because the trigonometric quantities are simply angular transforms of the depth variable.

When fitting k cycles to a log segment, the resulting wave form is known as the k th harmonic of the data. A complete set of harmonics is known as a Fourier series, which, when summed together, recreates the original log in its entirety:

$$Y = \sum A \cos k\theta + B \sin k\theta$$

The summation range extends between zero and a harmonic number of $n/2$, where n is the number of data points in the sequence. This highest harmonic corresponds to the Nyquist frequency. This is a limit beyond which higher frequency wavelets cannot be estimated and is set by the data spacing. Because the harmonic terms are orthogonal, the series represents a set of sine waves that are uncorrelated with one another. The series is conventionally shown by a periodogram that graphs the power (variance) against the harmonic number.. Alternatively, the harmonic numbers may be represented by a scale of their corresponding wavelengths, and the wave amplitude plotted instead of the power (the square of the amplitude).

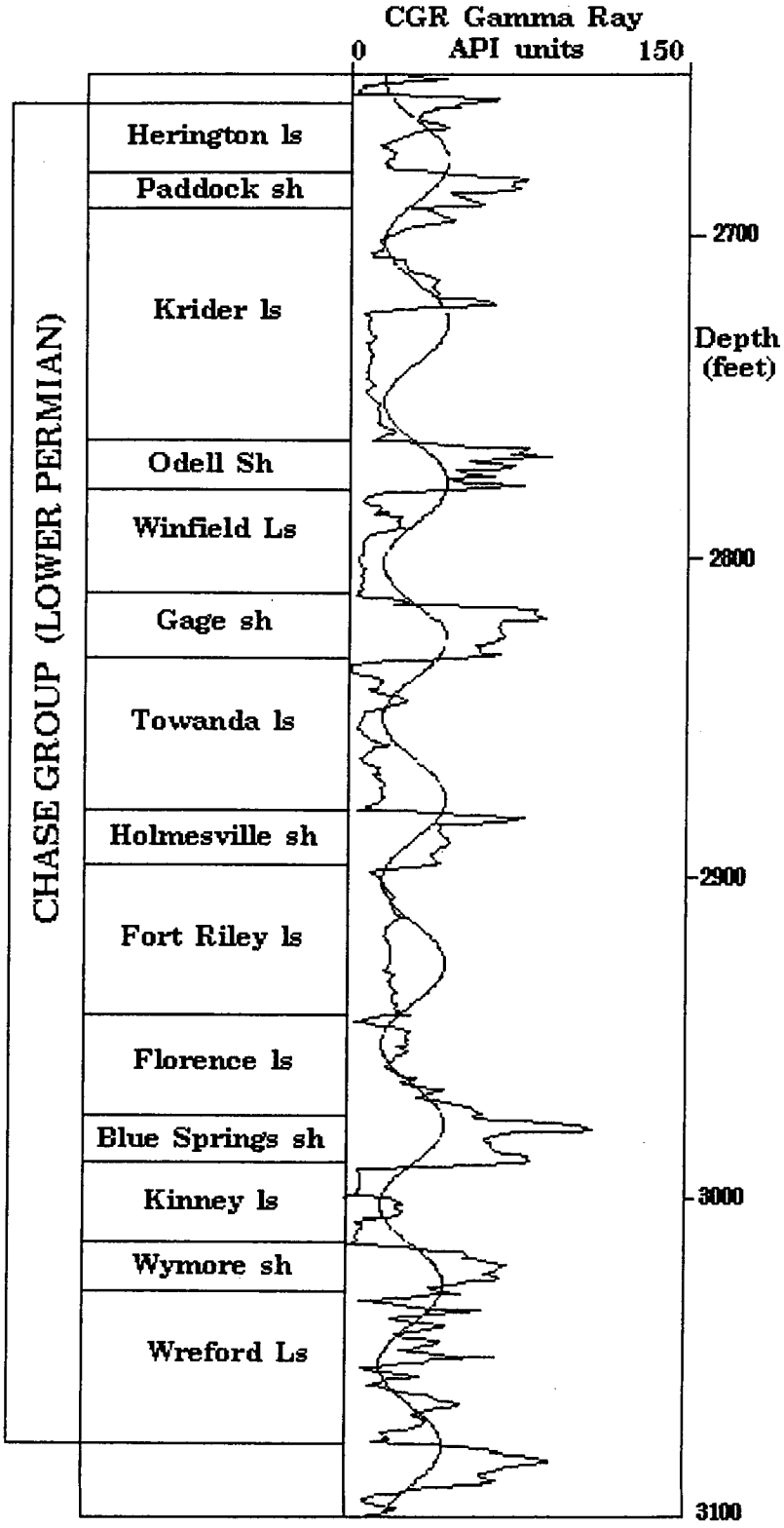


Periodogram of a gamma-ray log of the Lower Permian Chase Group: (a) original log, (b) log minus mean, (c) log first derivative.

This is an introductory treatment of the mathematics that underlie basic and discrete Fourier analysis. The concepts are best understood by reference to a real example. The Lower Permian Chase Group is a good subject for Fourier analysis, because its component lithologies have been thought for many years to be ordered in cyclothems. The group consists of a repetitive sequence of marine carbonates and supratidal shales. In southwestern Kansas, the generalized setting was one of tidal flats and shallow marine environments. The lithological record presumably reflects the operation of oscillatory processes on the carbonate shelf, driven by tectonic, eustatic, or more localized phenomena. Fourier analysis can be used to isolate any systematic cyclic signal and to estimate its wavelength and stratigraphic position. The computed gamma-ray log is primarily a measure of shale content and should be a useful record of the alternations between prograding clastic sources to the northwest and marine carbonate transgressions from the southeast.

Three alternative periodograms of the gamma-ray log are shown. The periodogram at the top gives the power or variance associated with each harmonic of the original log. The spectrum is dominated by the zero harmonic. This harmonic represents a wave with an infinite period, and so is a constant value positioned at the mean gamma-ray value. The power of the zero harmonic thus simply reflects the mean value. If a periodogram is computed for the log after the mean has been subtracted, then the zero harmonic is eliminated from the resulting periodogram. The spectrum shows a pronounced peak at the ninth harmonic, which corresponds to a wavelength of 50 ft. In computing a periodogram it is customary to remove any long-term trend from the data in order to suppress broad changes and emphasize distinctive cyclic phenomena with higher frequencies. There are a variety of methods to do this that depend to some degree on the understanding of what constitutes a long-term trend. A polynomial regression will fit the raw variation with a lazy curve, acting as a high-pass filter that concentrates higher frequencies within the residuals. Both long- and intermediate-term fluctuations can be absorbed through the computation of the first derivative of the log by subtracting adjacent log values. The third periodogram graphs the power of harmonics computed from the log of the first derivative and strongly accentuates the higher frequency harmonics. The ragged and complex character in the high-frequency range is largely a reflection of the stochastic nature of gamma-ray measurements. Much of the character can be explained in terms of error noise, which could be dampened by some local smoothing of the log prior to the computation of this periodogram.

The conversion of the gamma-ray log into a periodogram is a transformation from the time domain into the frequency domain. Conversely, each harmonic represents a repeating and simple sine wave whose wavelength is given by the total length of the record divided by the harmonic number. The high proportion of the total variance picked up by the ninth harmonic means that the corresponding sine wave should be a good representation of the major pattern of shale–carbonate alternation. The representation by a single harmonic suggests the operation of a systematic mechanism to produce such a regular spacing. If the period was more irregular, then the spectrum would tend to be smeared over a range of frequencies.



Gamma-ray log of the Lower Permian Chase Group fitted with a ninth harmonic sine wave.

A single k th harmonic may be converted to its equivalent wave form in the time domain through the use of the Fourier equation:

$$Y = A \cos k\theta + B \sin k\theta$$

where θ is the angular conversion of the depth into k cycles and calculated by:

$$\theta = \frac{(D - D_0)2\pi}{k}$$

where D_0 is the depth of the top of the sequence marked as a zero angular reference point. The coefficients of A and B are taken from the Fourier series computation used to generate the periodogram. The sine wave of the ninth harmonic is shown superimposed on the gamma-ray log of the Chase Group section. As anticipated, the extremes of the sine wave show a good match with the shale and carbonate subdivisions.

The discrete Fourier series is widely used in time series analysis because so many records are measurements that are taken at intervals of a day or some other time unit. A continuous spectrum of component frequencies can be generated by the Fast Fourier Transform (FFT) introduced by Cooley and Tukey (1965). The computer requirements are greater than those used in the Fourier analysis of discrete harmonics. Basically, the Fourier transform is expressed in an exponential form that incorporates imaginary components (functions of the square root of minus one). The FFT then finds the complex coefficients of all wavelengths down to the Nyquist frequency. A limiting condition of the FFT operation is that it must be applied to a record whose number of sample points is a power of two.

The Fast Fourier Transform is used routinely in seismic processing by geophysicists, who find it convenient, or even preferable, to operate in the frequency domain. Most geologists feel more comfortable in the time domain of the stratigraphic framework. However, the use of spectral analysis by geologists has become increasingly common. In part, this is because notions of cyclicity in the rock record have a long history and geologists are making increasing use of the computer. In addition, recent models of climate forcing by cyclic astronomical phenomena have been a major stimulus to the search for systematic periodic components that can be linked with correlatable time events. The emergence of the concepts of Milankovitch cycles and sequence stratigraphy has also encouraged spectral analyses and cyclic interpretations.

PROPERTIES OF NONSINUSOIDAL CYCLES

Cyclical properties in time are most commonly modelled by a sine wave. The sine wave has useful mathematical properties which make the Fourier transform a particularly powerful means to view records such as logs in terms of their frequency content. However, the sine wave convention is most commonly accepted because it is seen to be the actual representation of physical phenomena. Light waves and other wave forms from the electromagnetic spectrum are often cited as classic examples of naturally occurring sinusoidal forms. Harmuth (1977) pointed out that this is not necessarily so. Prisms and diffraction gratings are merely devices that decompose light into a set of sine waves. The mathematics of the Fourier transform do the same operation. Regardless of whether or not a log actually has sinusoidal components, the transformation results in its representation by a series of uncorrelated sine waves spread over a range of frequencies.

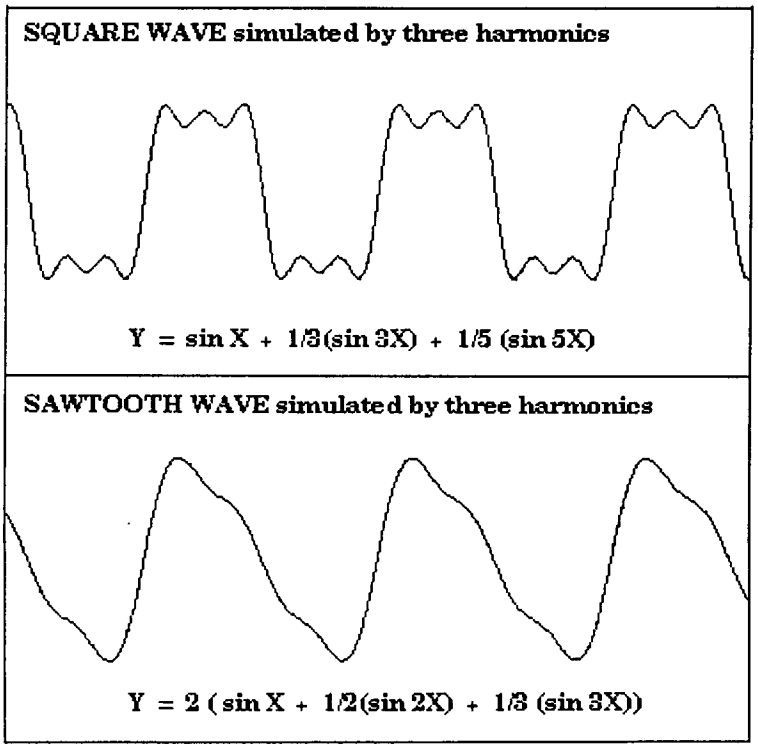
Examples of regular nonsinusoidal waves are common. A square wave can be fitted closely by the simple function of odd harmonics:

$$Y = \sin X + \frac{1}{3}(\sin 3X) + \frac{1}{5}(\sin 5X) + \ldots$$

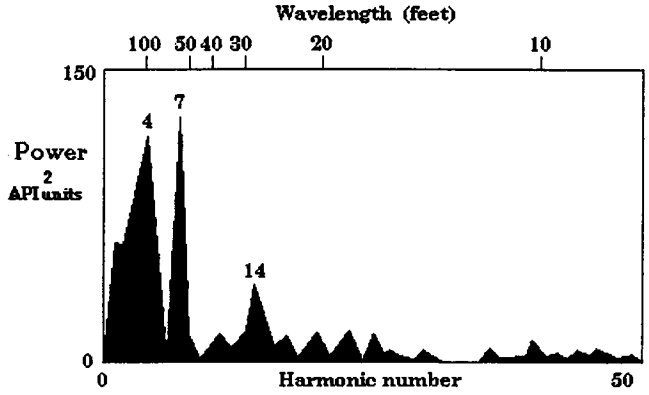
when taken to enough terms. A sawtooth wave can be built from both odd and even harmonics using:

$$Y = 2\left(\sin X + \frac{1}{2}\sin 2X + \frac{1}{3}\sin 3X + \ldots\right)$$

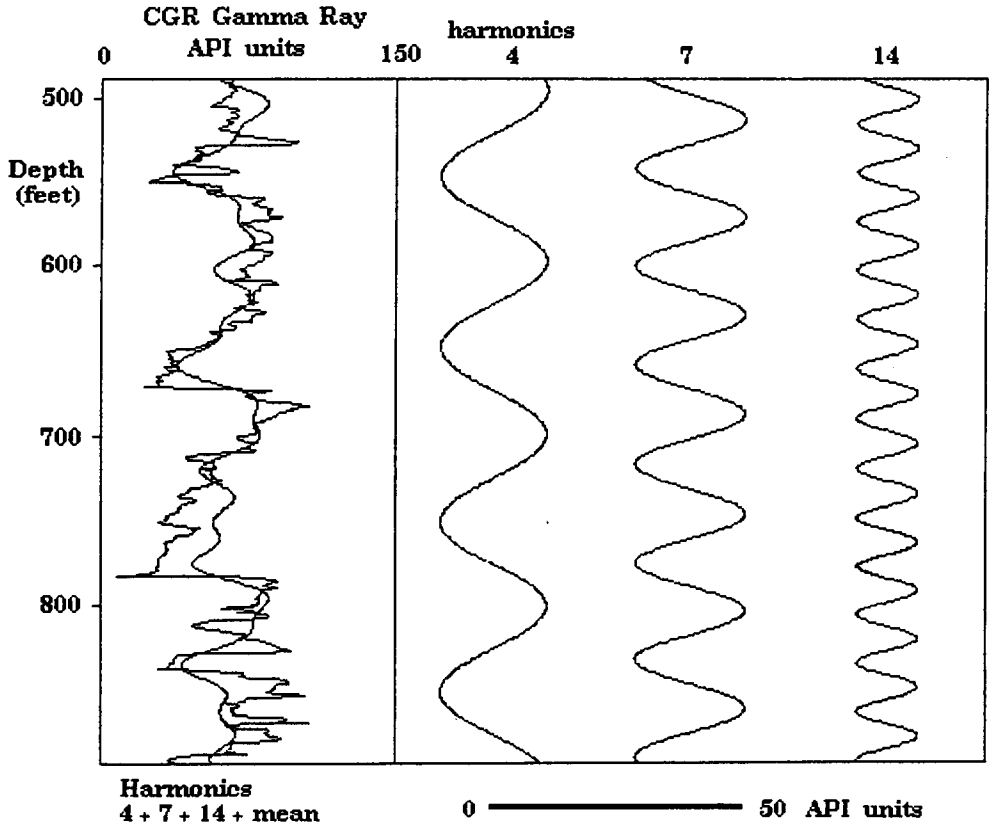
In the analysis of conventional time series, the presence of harmonics in an amplitude or power spectrum is most commonly interpreted to suggest that the fundamental cyclic pattern is nonsinusoidal (Chatfield, 1975).



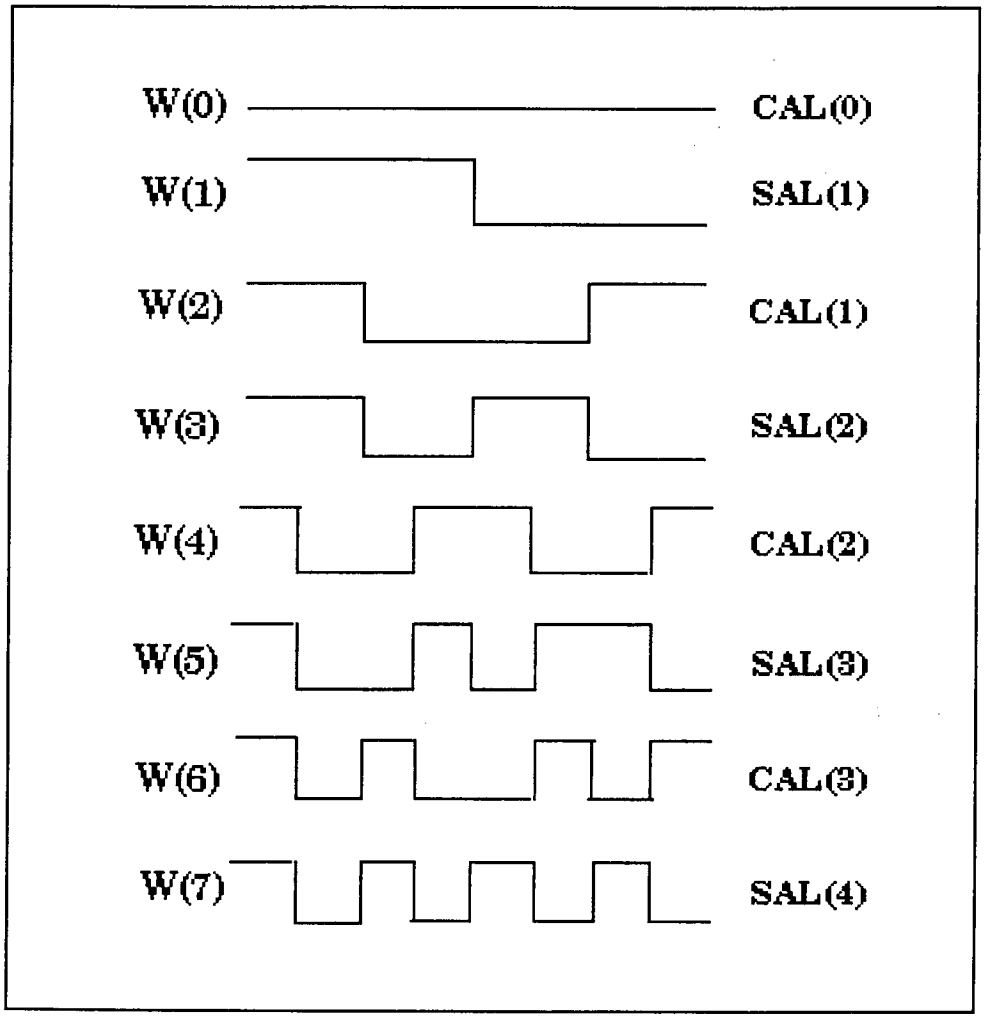
A practical illustration is given by the periodogram of the gamma-ray log from a Lower Cretaceous clastic succession in central Kansas. The periodogram is marked by three strong peaks at the fourth, seventh, and fourteenth harmonics.



The series is eerily close to simple integer multiples of a fundamental harmonic. Using the methods described earlier, the harmonics can be transformed back to the time domain. When compounded, they show a good generalized match with the original gamma-ray log. The harmonic pattern appears to have been triggered by the sawtooth motif of the log profile. The shapes are consistent with the origin of the thick sandstones in channels, with sharp erosional bases and fining-upward profiles. While there seems to be a fundamental repetition with a wavelength of about 100 ft, the multiple harmonics are more likely to be caused by a nonsinusoidal character, rather than additional cyclic processes.

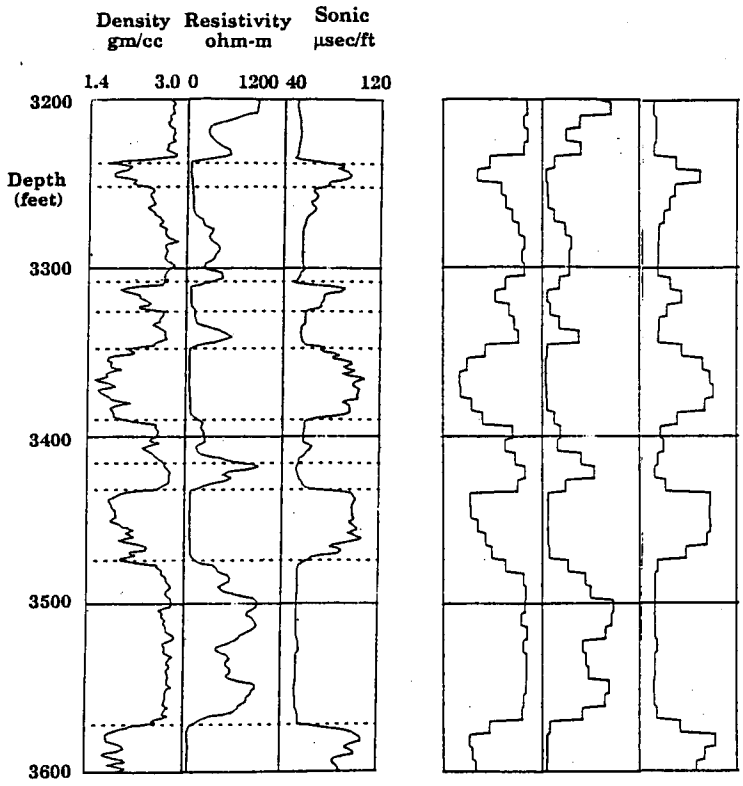


A considerable body of research and applications have been made using square waves in recent years. Although several alternative systems of rectangular wave forms have been proposed, the most widely used are square waves that take the value of either $+1$ or -1. A hierarchy of independent square waves is described by a set of Walsh functions. The concepts and terminology of Walsh functions have much in common with the Fourier series. Instead of frequency, a Walsh function is characterized by a sequency, which is the number of zero crossings per time unit divided by two. The even-numbered functions are symmetric and are called CAL functions because they are analogous to cosine waves. The odd-numbered, asymmetric members are SAL functions, squared versions of sine waves. The mathematics of Walsh functions are simpler than the equivalent trigonometric Fourier operations, so that a Fast Walsh Transform (FWT) runs much faster than a Fast Fourier Transform.



A conversion to a set of Walsh functions is a transformation to the sequency domain. As with the Fourier transform, the original signal can be recreated exactly from its Walsh transform. Alternatively, low-pass, high-pass, or band-width versions of the signal can be generated by selective summation of Walsh functions of appropriate sequency. Lanning and Johnson (1983) used this property as a means to block well logs in a consistent manner. By excluding

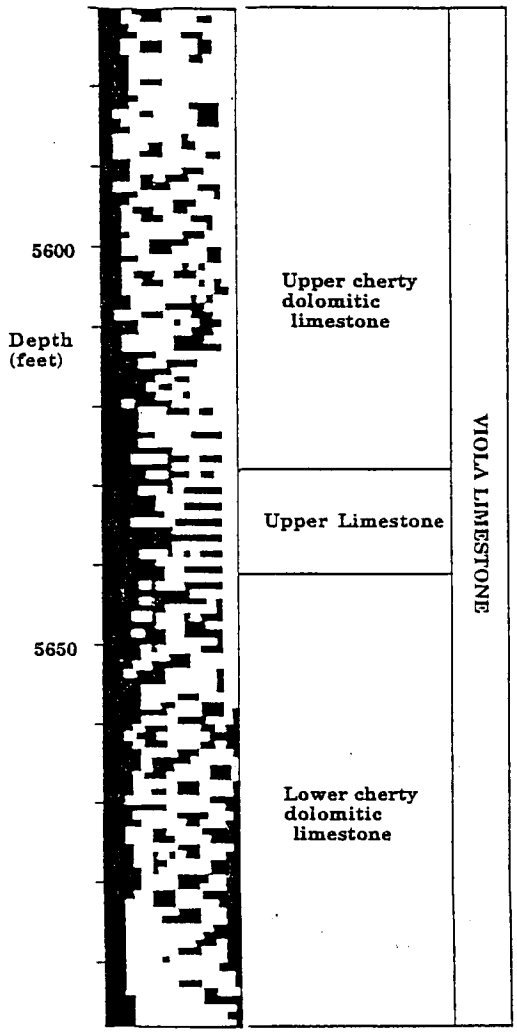
Walsh functions with sequencies that exceeded a "minimum resolvable layer thickness," they used the Walsh transform as a low-pass procedure to block well logs as stepped functions.



The application of Walsh functions to logging traces is reasonable since many rock sequences are made up of beds that are fairly homogeneous, but differentiated from one another by sharp boundaries. The basic logging concept of zones is a model that seeks to recover a stepped sequence of layers from the continuous log trace. Ideally, the result will emulate a succession of beds seen in an equivalent outcrop. The discrete nature of many bedding sequences has encouraged the use of Walsh spectra in preference to Fourier transforms (Weedon, 1989). In addition, Walsh spectra are less sensitive to abrupt changes at hiatuses, which tend to generate high frequency Fourier components. By the same token, the abrupt ramping of the lower sequency Walsh functions may cause them to be poor descriptors of longer term gradual changes.

Ultimately, the choice of Walsh or Fourier transform should be made with regard to the nature of the process model. Each transform type has its strengths and limitations. Just as the Fourier transform makes an overly complex description of a blocked function, the Walsh transform requires multiple square waves to approximate a sine wave. In addition, the representation in the sequency domain is not time invariant, as is the case with the frequency domain. Shifts in the starting point of the series can cause noticeable changes in the Walsh spectrogram. However, this shortcoming can be circumvented by producing an averaged Walsh spectrum (Beauchamp, 1984). Clearly, some intelligent choices should be made concerning the purpose of the analysis and the potential form of the signal at the scale of interest.

Doveton (1986) provided an example of a Walsh transform computed for a short sequence segment. The computations are very simple and form the basis for a sliding window filter program (Ahmed et al., 1976) to generate a moving Walsh spectrogram. This spectrogram graphs sequency amplitudes as a function of depth. Sequency increases from left to right and the spectrum is clipped to differentiate high amplitudes (black) from low amplitudes (white). Essentially, the spectrogram shows changes in the scales of log activity with depth. In this Viola Limestone example, the log records calcite content computed from a compositional analysis based on neutron, density, and sonic logs. Among other things, the calcite log gives indications of the relative thickness of limestone beds within the Viola Limestone section. The spectrogram shows how the patterns of thickness distribution change with stratigraphic level. Notice, for example, how the lower cherty dolomitic limestone is marked particularly by high-sequency (thin) features, and is contrasted with the low- and intermediate-sequency character of the upper limestone. When applied to a gamma-ray log, the spectrogram typically shows shale interbedding characteristics. However, the same technique can be used to analyze reservoir structure through the transformation of a porosity log.



Moving Walsh spectrogram of calcite variation in a Viola Limestone section. Sequency increases from left to right. Modified from Doveton (1986).

MARKOV CHAIN ANALYSIS

There are many situations in which a sequence of events in either time or space is observed as a succession of states taken from a limited set of alternatives. Petroleum geology and engineering are concerned with rock successions where observations spaced at constant intervals record the occurrence of lithology types or flow units. The relationship between adjacent events may be summarized by a transition tally matrix in which each cell sums the number of times that one state (identified by the row) is succeeded by another (identified by the column). So, for example, a succession consisting of lithologies A, B, C, and D may be summarized as *transition tally matrix* by taking observations at successive one-foot intervals, accumulating transition totals in the appropriate cells and might appear as:

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 8 & 4 & 3 & 1 \\ 4 & 5 & 2 & 1 \\ 1 & 2 & 4 & 5 \\ 3 & 1 & 3 & 3 \end{bmatrix} \end{array}$$

where for example, the number of times C succeeds A is 3. It can be seen that the *i*th row total is equal to the *i*th column total, which is a property of all transition matrices of this type, since every lithology that is entered is also left (with the exception of the initial and terminal events). The row (or column totals) may be written as the vector:

$$\begin{bmatrix} 16 & 12 & 12 & 10 \end{bmatrix}$$

which represents the number of times that the succession observations are in each of the four states.

Division of the tally matrix by each of the row totals leads to a *transition probability matrix*, **P**:

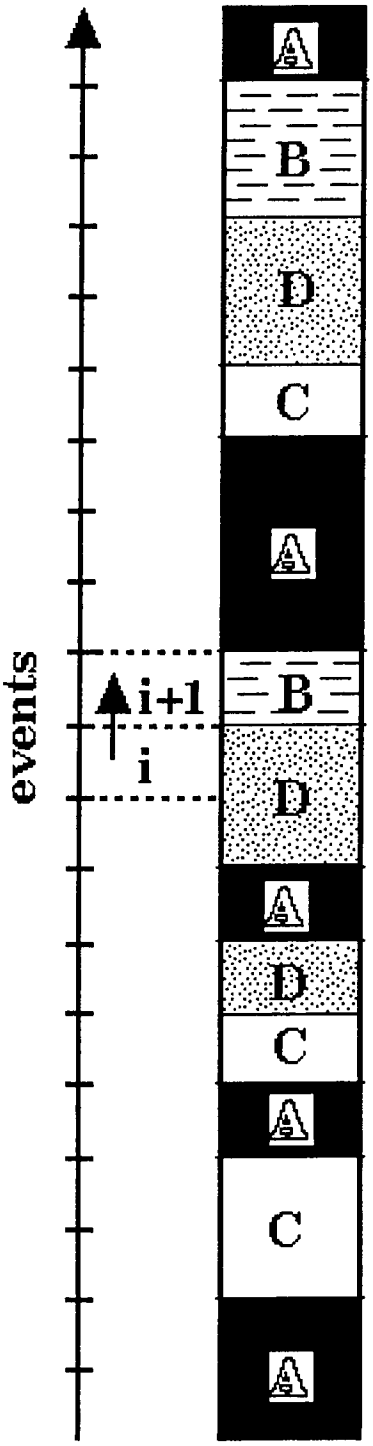
$$\underline{\mathbf{P}} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.50 & 0.25 & 0.19 & 0.06 \\ 0.33 & 0.42 & 0.17 & 0.08 \\ 0.08 & 0.17 & 0.33 & 0.42 \\ 0.30 & 0.10 & 0.30 & 0.30 \end{bmatrix} \end{array}$$

Similarly, division of the totals vector by the grand total results in an estimate of the *fixed probability vector*:

$$\begin{bmatrix} 0.32 & 0.24 & 0.24 & 0.20 \end{bmatrix}$$

which expresses the proportions of each lithology in the entire sequence.

These basic concepts are demonstrated by working through the simple illustrative example shown in the figure. The counting of transitions in the short sequence of events is tabulated as the transition tally matrix. The tally matrix is then converted to the transition probability matrix and fixed probability vector by the same operations described in the preceding text.



	A	B	C	D
A	3	-	3	1
B	2	1	-	-
C	1	-	1	2
D	1	2	-	2

Transition tally matrix

	A	B	C	D
A	0.43	0.00	0.43	0.14
B	0.67	0.33	0.00	0.00
C	0.25	0.00	0.25	0.50
D	0.20	0.40	0.00	0.40

Transition probability matrix

A	B	C	D
0.37	0.16	0.21	0.26

Fixed probability vector

Because A, B, C, and D are mutually exclusive events, the probability that one state is followed by another is either a conditional or an unconditional probability. P(B/ A) is the notation that A will be followed by B, given that A has occurred as the previous event. In the unconditional case:

$$P(B_{i+1} / A_i) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A)\ P(B)}{P(A)} = P(B)$$

as opposed to the conditional alternative where:

$$P(B / A) \neq P(B)$$

If all transitions are unconditional, then:

$$P(B / A) = P(B / B) = P(B / C) = P(B / D)$$

and the model is one of *independent events*. The expected transition probability matrix, **A**, for independent events consists of rows of the fixed probability vector. For the example that we started with:

$$\underline{\mathbf{A}} = \begin{bmatrix} 0.32 & 0.24 & 0.24 & 0.20 \\ 0.32 & 0.24 & 0.24 & 0.20 \\ 0.32 & 0.24 & 0.24 & 0.20 \\ 0.32 & 0.24 & 0.24 & 0.20 \end{bmatrix}$$

A matrix of expected tallies can be computed by multiplying the probabilities by the row totals of the observed tally matrix. The null hypothesis of independent events may be tested as a chi-square contingency with $(m\text{-}1)^2$ degrees of freedom. If the null hypothesis is rejected, then the alternative model is accepted of a partial dependency between adjacent events and is known as a *finite Markov chain*. When the sequential pattern is controlled entirely by relationships between adjacent events, then the chain is of first order.

If the sequence is shown to have a Markov property, then the structural form of the Markov chain may be deduced from a comparison of the observed transition frequencies with those expected from the independent events predictions. Transition types that occur more frequently than expected can be linked together as a *preferred transition path* that is a graphic description of the chain. In geological applications, the pattern of preferred transitions reflect depositional changes that are elements of sedimentary facies models. The pattern can also be interpreted readily in terms of repetitive motifs such as cycles or rhythms because of the limited number of states in the Markov chain.

Named after its discoverer A.A. Markov, whose inspiration was the alternation of vowels and consonants in Pushkin's poem "Onegin", Markov chain models are an example of a stochastic process model. Markov chain models occur in the range between the extremes of determinism, where every event is exactly

specified by its predecessor, and independent events, where there is no relationship between successive events.

If the matrix P is squared:

$$\mathbf{P}^2 = \mathbf{P}.\mathbf{P}$$

the resulting matrix is the expectation of the probability of the (i+2) event given that of the ith event, as predicted by the first-order Markov chain. If the matrix differs significantly from that observed in the sequence (as judged by a chi-square test), then the sequence has second order Markov properties. In general, if the matrix **P** is successively powered to the limit, the matrix converges to a matrix of equilibrium proportions of the states, whose rows match the fixed probability vector.

When a sequence is shown to have a first order Markov property then the transition probability matrix is a significant descriptor of the transition properties. However, the description is constrained to the relationship between immediately successive events and independent of all prior events. The type of dependency is often known as the *memory* of the process, which in this instance, has a length of one step (the distance between adjacent events). In a second-order Markov model, the state event immediately preceding the pair of adjacent events is incorporated as part of the conditional probability. The memory contained within the transition probabilities is now extended to two steps.

The concept can be generalized to nth-order Markov models with memories of n steps. At one extreme, a memory of no steps is an independent -events situation where past events have no influence on the present event and the succession is random. Models with long memories suggest complex patterns of ordering with systematic, long-term components that are relatively free from short-term disruptions. In practice, many rock sequences can be represented adequately by low-order Markov models (Doveton, 1971). A first-order Markov model is often a sufficiently useful representation for descriptive purposes. However, the simulation of entire successions from low-order transition probability matrices involves scaling-up considerations, because the model is controlled by a short-term memory. If there are systematic longer term mechanisms, they will not be captured by the transition probabilities. This situation commonly occurs when the succession is markedly *non-stationary* because of long-term trends. The transition probabilities will then change in value over the length of the succession. However, the problem can be accommodated by using several sets of transition probabilities keyed to sub-sequences or by linking the probabilities with position within the succession.

Some authors also experimented with the modeling of sedimentary successions from transition probability matrices. Simulation examples ranged in scale from the internal structure of fluvial sandstones (Potter and Blakely, 1967), through simulation of the Pennsylvanian rock sequence (Schwarzacher, 1967), the lateral migration of transgressive-regressive strand-line deposits (Krumbein, 1968), to marine sedimentation in space and time (Harbaugh, 1966). At the time, these Markov simulation studies had academic interest because of the limited use of computers in geology, but are now one of the methods used routinely to model reservoirs, particularly with the emergence of stochastic reservoir models (Haldorsen and Lake, 1984) .

MARKOV CHAIN ANALYSIS AND SIMULATION OF SEDIMENTARY SEQUENCES IN ONE DIMENSION

The fundamental Markovian descriptor of a sequence is the transition probability matrix, P, which takes the form:

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

for a first-order Markov chain with (in this example) three states. In general applications, observations of the state of a process are taken at equal increments of time. The time unit for each step is chosen so as to be small enough so that all events of significance are recorded in the sequence of transitions.

The selection of the length of the interval between successive events must be considered carefully. If too small an interval is chosen, then the number of transitions of states to themselves becomes extremely large. Statistical tests for a significant Markov property will then reveal the trivial fact that successive observations tend to be repetitions of the same state. If too large an interval is used, many thin bed events are missed altogether and the chain is biased towards states that tend to have thicker beds.

The off-diagonal elements of the transition probability matrix capture the transition characteristics between states and will be the same for any sampling interval finer than the thinnest bed. The transition probabilities of a state to itself occur on the main diagonal of the transition probability matrix. They dictate the statistics of the distribution of thicknesses of each state and will vary with the length of the interval used. The mean thickness, m, and its variance, v, of the state i are easily computed from the equations:

$$m = \frac{1}{(1 - p_{ii})}$$

and

$$v = \frac{p_{ii}}{(1 - p_{ii})^2}$$

where the units are in number of interval spacings (Kemeny and Snell, 1960). The equations of these parameters show that thicknesses implied by Markov transition probabilities follow a geometric distribution (Krumbein and Dacey, 1969). This becomes an important consideration if the Markov transition probability matrix is used to model a synthetic stratigraphic succession. Studies of sedimentary bed thicknesses have generally concluded that they are lognormally distributed (see e.g. Pettijohn, 1957; Potter and Siever, 1955). However, this identification comes from the empirical observation that bed thicknesses are fitted approximately by straight lines when plotted on a log probability grid. A theoretical model of bed sedimentation based on simple probability will generate a geometric thickness distribution as pointed out by Krumbein and Dacey (1969). Of course, such a model is intrinsically Markovian.

MARKOV CHAIN SIMULATION ON AN EXCEL SPREADSHEET

MARKOV CHAIN SEQUENCE SIMULATOR

Initial state (A, B, or C) = B

Transition probability matrix

P=	A	B	C
A	0.8	0.2	0
B	0.1	0.7	0.2
C	0.1	0.3	0.6

Lookup table

	A	B	C	Z
A	0	0.8	1	1
B	0	0.1	0.8	1
C	0	0.1	0.4	1

t	X(t)	RAND	X(t+1)					Depth	
0	B	0.812	C	0	0.1	0.8	1	1	A
1	C	0.082	A	0	0.1	0.4	1	2	A
2	A	0.939	B	0	0.8	1	1	3	A
3	B	0.667	B	0	0.1	0.8	1	4	B
4	B	0.583	B	0	0.1	0.8	1	5	B
5	B	0.187	B	0	0.1	0.8	1	6	C
6	B	0.185	B	0	0.1	0.8	1	7	B
7	B	0.224	B	0	0.1	0.8	1	8	B
8	B	0.373	B	0	0.1	0.8	1	9	B
9	B	0.332	B	0	0.1	0.8	1	10	A
10	B	0.883	C	0	0.1	0.8	1	11	C
11	C	0.558	C	0	0.1	0.4	1	12	C
12	C	0.674	C	0	0.1	0.4	1	13	C
13	C	0.773	C	0	0.1	0.4	1	14	C
14	C	0.362	B	0	0.1	0.4	1	15	C
15	B	0.269	B	0	0.1	0.8	1	16	B
16	B	0.668	B	0	0.1	0.8	1	17	B
17	B	0.899	C	0	0.1	0.8	1	18	B
18	C	0.298	B	0	0.1	0.4	1	19	B
19	B	0.085	A	0	0.1	0.8	1	20	B
20	A	0.499	A	0	0.8	1	1	21	B
21	A	0.838	B	0	0.8	1	1	22	B
22	B	0.56	B	0	0.1	0.8	1	23	B
23	B	0.266	B	0	0.1	0.8	1	24	B
24	B	0.209	B	0	0.1	0.8	1	25	B
25	B	0.346	B	0	0.1	0.8	1	26	B
26	B	0.112	B	0	0.1	0.8	1	27	B
27	B	0.155	B	0	0.1	0.8	1	28	B
28	B	0.332	B	0	0.1	0.8	1	29	B
29	B	0.836	C	0	0.1	0.8	1	30	B
30	C	0.017	A	0	0.1	0.4	1	31	B
31	A	0.866	B	0	0.8	1	1	32	B
32	B	0.657	B	0	0.1	0.8	1	33	B
33	B	0.823	C	0	0.1	0.8	1	34	C
34	C	0.061	A	0	0.1	0.4	1	35	C
35	A	0.544	A	0	0.8	1	1	36	B
36	A	0.806	B	0	0.8	1	1	37	B
37	B	0.707	B	0	0.1	0.8	1	38	B
38	B	0.504	B	0	0.1	0.8	1	39	B
39	B	0.49	B	0	0.1	0.8	1	40	A
40	B	0.774	B	0	0.1	0.8	1	41	C
41	B	0.249	B	0	0.1	0.8	1	42	C
42	B	0.874	C	0	0.1	0.8	1	43	C
43	C	0.171	B	0	0.1	0.4	1	44	C
44	B	0.433	B	0	0.1	0.8	1	45	C
45	B	0.55	B	0	0.1	0.8	1	46	C
46	B	0.8	B	0	0.1	0.8	1	47	C
47	B	0.165	B	0	0.1	0.8	1	48	B
48	B	0.618	B	0	0.1	0.8	1	49	B
49	B	0.491	B	0	0.1	0.8	1	50	B
50	B	0.393	B	0	0.1	0.8	1	51	B
51	B	0.228	B	0	0.1	0.8	1	52	B
52	B	0.443	B	0	0.1	0.8	1	53	B
53	B	0.822	C	0	0.1	0.8	1	54	B
54	C	0.535	C	0	0.1	0.4	1	55	B
55	C	0.772	C	0	0.1	0.4	1	56	B
56	C	0.43	C	0	0.1	0.4	1	57	B
57	C	0.602	C	0	0.1	0.4	1	58	C
58	C	0.456	C	0	0.1	0.4	1	59	B
59	C	0.772	C	0	0.1	0.4	1	60	B
60	C	0.099	A	0	0.1	0.4	1	61	B
61	A	0.859	B	0	0.8	1	1	62	B
62	B	0.405	B	0	0.1	0.8	1	63	B
63	B	0.564	B	0	0.1	0.8	1	64	B
64	B	0.697	B	0	0.1	0.8	1	65	A
65	B	0.941	C	0	0.1	0.8	1	66	A
66	C	0.562	C	0	0.1	0.4	1	67	C
67	C	0.298	B	0	0.1	0.4	1	68	B
68	B	0.638	B	0	0.1	0.8	1	69	B
69	B	0.244	B	0	0.1	0.8	1	70	A
70	B	0.256	B	0	0.1	0.8	1	71	C
71	B	0.648	B	0	0.1	0.8	1	72	B
72	B	0.537	B	0	0.1	0.8	1	73	B
73	B	0.181	B	0	0.1	0.8	1	74	B
74	B	0.619	B	0	0.1	0.8	1	75	B
75	B	0.132	B	0	0.1	0.8	1	76	B
76	B	0.132	B	0	0.1	0.8	1	77	B
77	B	0.362	B	0	0.1	0.8	1	78	B
78	B	0.583	B	0	0.1	0.8	1	79	B
79	B	0.735	B	0	0.1	0.8	1	80	A
80	B	0.467	B	0	0.1	0.8	1	81	A
81	B	0.451	B	0	0.1	0.8	1	82	B
82	B	0.539	B	0	0.1	0.8	1	83	C
83	B	0.442	B	0	0.1	0.8	1	84	B
84	B	0.674	B	0	0.1	0.8	1	85	B
85	B	0.961	C	0	0.1	0.8	1	86	B
86	C	0.481	C	0	0.1	0.4	1	87	C
87	C	0.543	C	0	0.1	0.4	1	88	C
88	C	0.984	C	0	0.1	0.4	1	89	C
89	C	0.421	C	0	0.1	0.4	1	90	C
90	C	0.077	A	0	0.1	0.4	1	91	B
91	A	0.99	B	0	0.8	1	1	92	B
92	B	0.373	B	0	0.1	0.8	1	93	B
93	B	0.706	B	0	0.1	0.8	1	94	B
94	B	0.986	C	0	0.1	0.8	1	95	B
95	C	0.124	B	0	0.1	0.4	1	96	B
96	B	0.389	B	0	0.1	0.8	1	97	B
97	B	0.001	A	0	0.1	0.8	1	98	B
98	A	0.654	A	0	0.8	1	1	99	A
99	A	0.162	A	0	0.8	1	1	100	C

If a transition probability matrix is used to model a synthetic succession, then systematic comparisons should be made with the real succession to verify that desired features are replicated by the simulation. The transition matrix of a first-order chain represents a very myopic view of the succession because it is restricted to the conditional relationships of adjacent events. Checks on long-term statistics will confirm whether the Markov simulation is a satisfactory match. If not, then a higher-order Markov chain may be appropriate or additional longer-term elements should be incorporated to supplement the Markov model.

The real succession should first be analyzed for stationarity. Do the transition probabilities stay effectively constant at all positions in the succession? If they do not, then the succession may have to be subdivided into segments that are effectively stationary. Alternatively, the transition probabilities can be modified to incorporate a drift component that reflects relative depth position. Is a first-order Markov chain an adequate description of the transition properties of the sequence? A one-step memory may be insufficient to capture systematic transitional behavior, particularly if there are distinctive sequential patterns that pick up phenomena such as fining-upward or coarsening-upward trends. These alternatives can be checked by comparing first-order Markov predictions of two-step transition types with their actual frequencies (Doveton, 1971). If the first-order Markov prediction fails to be adequate, then the system can be expanded to a second-order (or higher) Markov model. So, for example, Schwarzacher (1967) elected to use a double-dependence Markov chain in his simulation of the Pennsylvanian sequence of Kansas.

Some of the longer-term properties of a succession will be honored automatically by the Markov simulation if the transitions are counted from the succession. At the most fundamental level, the proportions of the lithologies in the total succession are registered by the row and column totals of the transition tally matrix and these proportions are transferred into the transition probability matrix. In addition, Matalas (1967) pointed out that the first three moments of a succession will be preserved in a Markovian simulation.

The means and variances of the thicknesses of the lithologies in the succession should be compared with those calculated from the transition probability matrix. If they are radically different, then the actual dispersion of thicknesses appears to be represented poorly by a geometric distribution. The most common solution to this problem is to adopt a modified model, the *embedded Markov chain* (Gingerich, 1969). Rather than make observations at fixed intervals, the observation matrix is used to record only transitions between states. Transitions of a state to itself are now precluded and the main diagonal of the transition probability matrix has zero values. Each step of a simulation of a sequence based on this matrix now becomes a separate state. The thickness of each event can be generated by a random selection from the actual distribution of thicknesses of the corresponding state in the real succession.

Does the occurrence of the states throughout the synthetic succession show comparable characteristics to those observed in the real succession? Useful measures for this comparison are those of the *first-passage time statistics*.

(Doveton and Duff, 1984). The mean first-passage time expresses the average number of events that occurs after leaving a state before the same state is reentered. The variance measures the relative dispersion of passage time about the mean value. The expected passage time statistics can be calculated from the transition probability matrix as a matrix, **M**, of mean values, and **W**, of their variances. A fundamental matrix, **Z**, may be defined as:

$$\mathbf{Z} = (\mathbf{I} - \mathbf{P} + \mathbf{A})^{-1}$$

Then:

$$\mathbf{M} = (\mathbf{I} - \mathbf{Z} + \mathbf{E}\mathbf{Z}_d)\mathbf{D}$$

and:

$$\mathbf{W} = \mathbf{M}(2\mathbf{Z}_d\mathbf{D} - \mathbf{I}) + 2(\mathbf{Z}\mathbf{M} - \mathbf{E}(\mathbf{Z}\mathbf{M})_d)$$

where **E** is an *mxm* matrix of unit values ; $\mathbf{Z}_d$ is the diagonal matrix of **Z**; **D** is a diagonal matrix whose elements are the reciprocals of the fixed probability vector; and *m* is the number of states. Interested readers are referred to Kemeny and Snell (1960) for additional details on the derivation of these matrix algebra relationships.

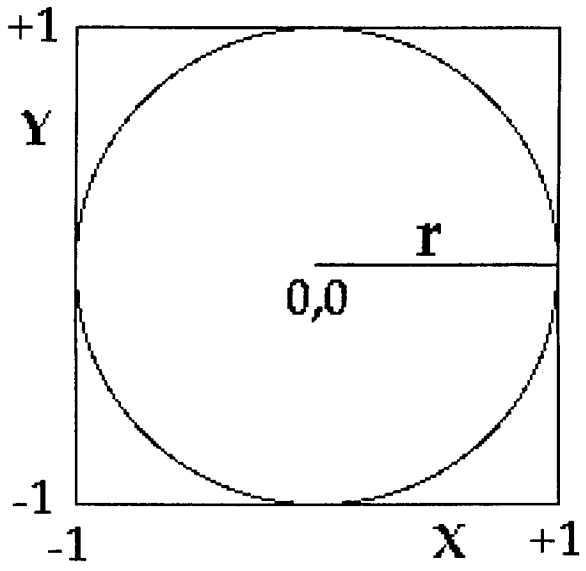
The passage time statistics are the Markovian prediction of the mean and variance of the number of lithologies that intervene between successive occurrences of any given lithological state. These measures are particularly useful when applied in successions believed to have been deposited as cyclic sequences. If one of the states can be identified with the initiation or conclusion of a cycle, then the state forms a reference marker and the passage time statistics of this state are Markovian descriptor of the cycle. Comparisons with actual cycles extracted from the real succession will verify whether the Markov simulation is adequate in their representation.

MONTE CARLO SIMULATION

The Monte Carlo method was introduced in 1946 by Stanislaw Ulam, a mathematician who worked on the Manhattan Project during World War II. He invented the Monte Carlo method in 1946 while pondering the probabilities of winning a card game of solitaire. The Monte Carlo method applies to any technique of statistical sampling used to approximate solutions to quantitative problems. Ulam did not invent statistical sampling, but pioneered the use of computers to automate the process. Statistical sampling had been applied by the early statisticians, but they used the outcomes of dice or card draws for their selection of a random sample. . Ulam worked with John von Neuman and Nicholas Metropolis to develop algorithms for computer implementations. Most importantly, they expanded the approach to also solve non-random problems that were too complex to solve by standard computational methods. Metropolis introduced the term Monte Carlo for the methodology with reference to the Casino. The first paper on the Monte Carlo method was published by Ulam and Metropolis in 1949.

Monte Carlo simulation is a stochastic method. The use of the term "stochastic" refers to the application of random numbers to generate an answer. Large numbers of repetitions of a Monte Carlo procedure create multiple scenarios of a problem which collectively converge on a solution. When repeated for many scenarios, the average solution will give an approximate answer. The accuracy of this answer can be improved by simulating more scenarios. For any given Monte Carlo simulation, the accuracy is proportional to the square root of the number of scenarios used. Monte Carlo simulation is a "brute force" approach that is able to solve problems for which no other solutions exist or where direct computation would be far more complex than the Monte Carlo equivalent.

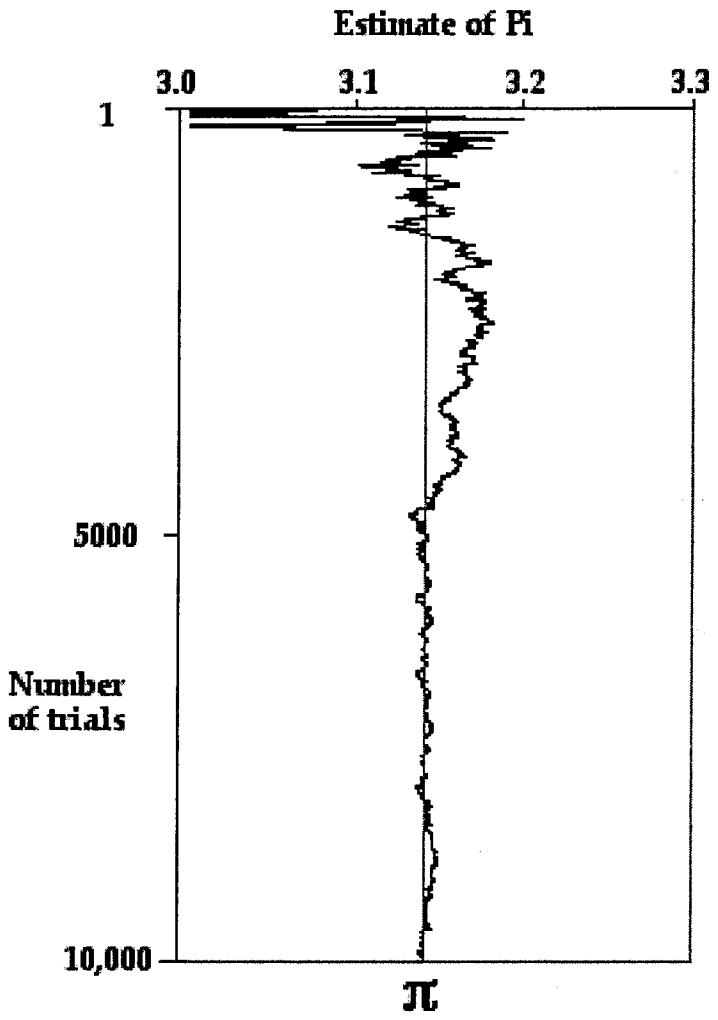
A simple demonstration of A Monte Carlo simulation would be a method aimed at solving the value of π. In the illustration below, a circle with radius r is precisely enclosed by a square, the length of whose sides is therefore 2*r.



Let us imagine that we throw thousands of darts randomly at the square. The proportion of darts that fall inside the circle ("hits") is given by the ratio of the area of the circle to the area of the square. The area of the circle is πr^2 and the area of the square is $4r^2$. Therefore the ratio and so proportion of hits is $\pi/4$.

We can design a Monte Carlo procedure to give us an approximate answer to the value of π, using these observations. In our simulation, we will have a circle whose radius is unity and with a center whose coordinates are (0,0). The enclosing square has reference axes of X and Y and the minimum and maximum values of X and Y in the square are -1 and +1. Now we will use a random number generator to determine the coordinates of the landing of our first dart. The EXCEL function RAND() generates random numbers between 0 and 1. By modifying the outcome to 2*RAND()-1, the random number will be between -1 and +1. We select two random numbers to specify the X and Y coordinates. Was the dart a "hit" (inside the circle) or a "miss" (outside the circle)? Now the distance of the dart from the center of the circle is the square root of the sum of the squares of X and Y (from the Pythagoras theorem). If the distance is less than 1, then the dart is inside the circle ("hit"). If we run the simulation repeatedly for the equivalent of thousands of darts and then divide the number of "hits" by the number of "darts" and multiply the result by 4, then this value will be an estimate of π whose accuracy will be basically a function of the square root of the number of iterations.

TRIAL	2*RAND()-1	2*RAND()-1	DISTANCE	HIT=1	SUMHITS	PI
1	-0.6271088	-0.4125081	0.75061865	1	1	4
2	-0.8197016	0.13324798	0.83046116	1	2	4
3	-0.9489369	0.91318233	1.31695985	0	2	2.66666667
4	0.78631912	0.86275786	1.16732553	0	2	2
5	-0.4323583	0.23224441	0.49078623	1	3	2.4
6	-0.7049662	0.95004699	1.18303278	0	3	2
7	0.83411242	0.24068603	0.86814359	1	4	2.28571429
8	0.20016281	0.22160782	0.29862213	1	5	2.5
9	0.83303832	-0.3080056	0.88815556	1	6	2.66666667
10	-0.5004452	-0.4501142	0.67308857	1	7	2.8
11	-0.1487196	-0.1525227	0.21302747	1	8	2.90909091
12	0.496872	0.20255787	0.53657383	1	9	3
13	-0.2565154	-0.8147366	0.8541638	1	10	3.07692308
14	0.89484149	0.66089869	1.11244253	0	10	2.85714286
15	-0.8912662	-0.7024869	1.13483183	0	10	2.66666667
16	-0.7014089	-0.1141054	0.71062964	1	11	2.75
17	-0.2068563	0.88692569	0.91072867	1	12	2.82352941
18	0.91981787	-0.0460885	0.9209718	1	13	2.88888889
19	-0.2122304	0.10770654	0.23799674	1	14	2.94736842
20	0.20856528	0.74230303	0.77104687	1	15	3
21	0.58074444	-0.0861662	0.58710196	1	16	3.04761905
22	0.18480453	-0.6121024	0.63939194	1	17	3.09090909
23	0.96503973	0.07780313	0.96817096	1	18	3.13043478



When the results of the trials are graphed, we can see the convergence of the estimate as an approximation of π. In this example run, the estimate after 10,000 trials was 3.1404. A larger number of trials would give a closer estimate, but remember that the accuracy is controlled by the square root of the number of trials, not the number of trials. Another issue is the difficulty of generating truly random numbers by an algorithm which is inevitably a non-random procedure. So, a random number generator is really a pseudo-random number generator with hopefully only minute differences from randomness. However, these minute and systematic differences may limit the solution to being an approximation, regardless of the number of trials.

MONTE CARLO SIMULATION EXAMPLE: FORWARD_MODELLING LOGS FROM SIMULATED RESERVOIR SECTIONS

Introduction

Conventional log analysis applies the Archie equation to wireline resistivity and porosity measurements in the computation of a water saturations in potential reservoir sections. The process is commonly (and maybe appropriately) termed "log interpretation", both because of the inherent uncertainties regarding the true values of the parameters in the Archie equation and the need to reconcile the calculated fluid distribution with the physics of a buoyant hydrocarbon column in its penetration of the pore structure of successive sedimentary units in a sandstone reservoir. The reasoning is an inverse approach in which log responses are used to deduce the rock properties that control them. Alternatively, "forward-modeling" can be applied to predict log responses, based on rock properties and capillary pressure measurements, and a comparison made between predictions and log curves in a search for the optimal reservoir model.

Forward-modeling software can be written as an EXCEL spreadsheet, using standard functions rather than specialized macros, as demonstrated by this report and the workbook on the attached diskette. The procedure follows the following steps (see Fig. 1):
(1) the identification of petrofacies to characterize rock units in terms of common associations of porosity, permeability, and fluid saturation:
(2) the construction of a transition probability matrix of the petrofacies to be used in creating a stratigraphic column of the petrofacies sequence as a Markov chain realization;
(3) the assignation of a representative porosity to each zone of the model sequence according to its petrofacies;
(4) the estimation of the permeability of each zone, based on its petrofacies and porosity;
(5) the prediction of the water saturation of each zone, using Pittman equation calibrations with height above free-water level (the base of the simulated section),and the zone porosity and permeability;
(6) the generation of the resistivity log curve for the simulated section, as a forward-modeling application of the Archie equation;
(7) the display of a Pickett plot of resistivity and porosity as an expectation of a log pattern to be compared with observed crossplotted logs.

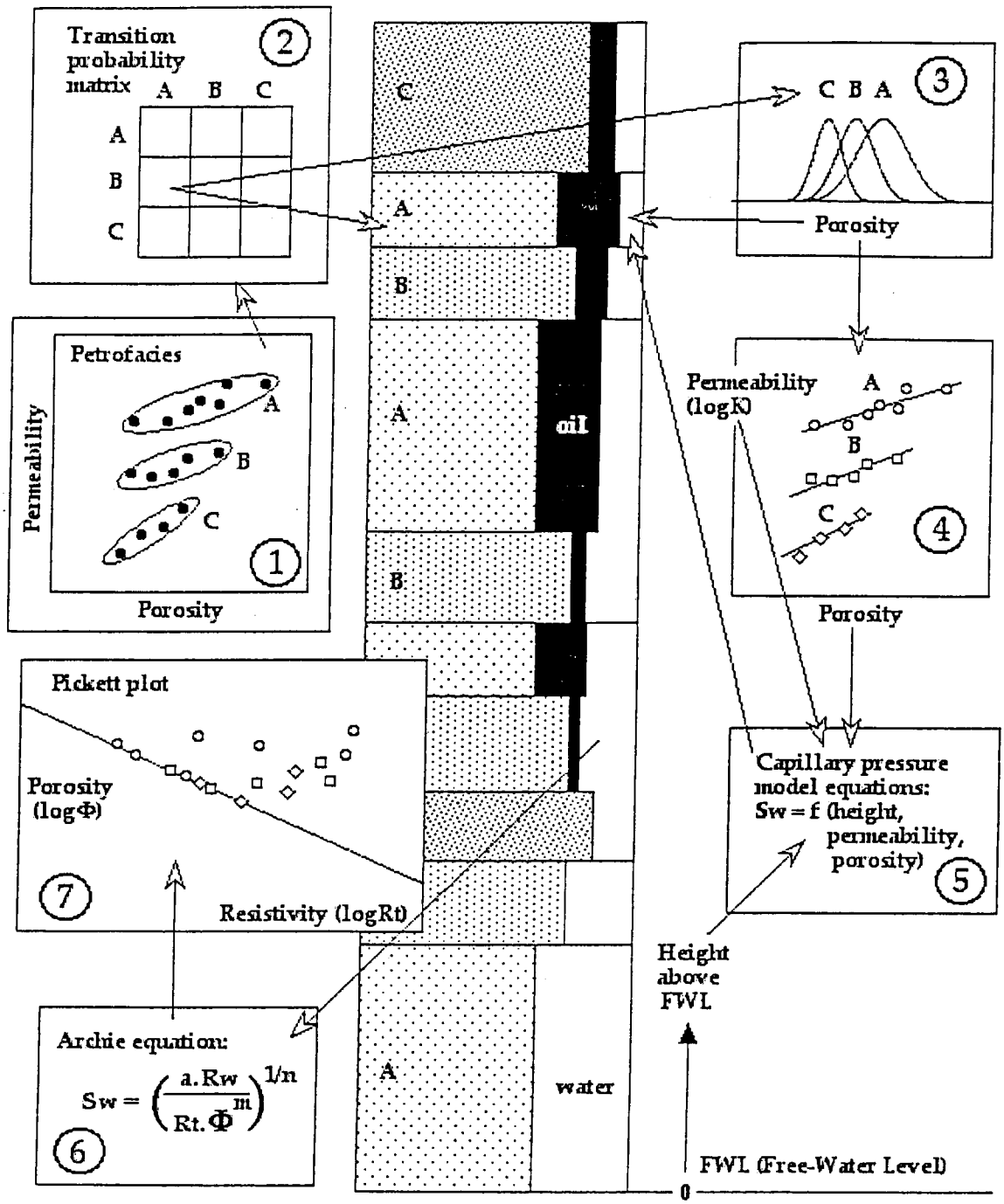


Figure 1. Flow diagram of operational steps (numbered in order) to simulate a sandstone reservoir and compute its wireline log responses.

1. Petrofacies identification

There are numerous ways to subdivide sandstones into facies, based on mineralogy, bedding structures, paleontology and other characteristics. However, the purpose of conventional geological facies is a classification that reflects the genesis of the sandstone and/or its diagenetic history. By contrast, effective petrofacies subdivision is morphological, rather than genetic, and should reflect the pore space of the rock in terms of volume, pore body and pore throat size distributions, geometry and connectivity. These factors control reservoir quality and differentiate flow units within a reservoir. Criteria to define petrofacies in sandstones may be based on petrographic observations of grain size, sorting, and other textural measures and/or petrophysical measures of porosity and permeability.

The example data set of this report used for demonstrating the modeling methodology is drawn from (unpublished) measurements of porosity, permeability, and grain-size of Simpson Sandstone core samples from Stafford County, Kansas by Alan Byrnes. Byrnes subdivided the sample set between medium-fine grained sandstone (petrofacies A), very fine-grained sandstone (petrofacies B), and very fine-grained/ silty sandstone (petrofacies C).The crossplot of porosity and permeabilty of these samples (Figure 2 and pfsim.xls, worksheet 'petrofacies') shows that the grain-size petrofacies subdivision is reflected by systematic and distinctive petrophysical trends, with increasing grain-size matched by higher permeability, as would be expected for the decrease in internal specific area. So, although the petrofacies were defined petrographically, the classification captures useful petrophysical distinctions between Simpson Sandstone poretype families. The measurement data and their statistical averages are shown in Figure 3 (and pfsim.xls, worksheet 'petrofacies'),

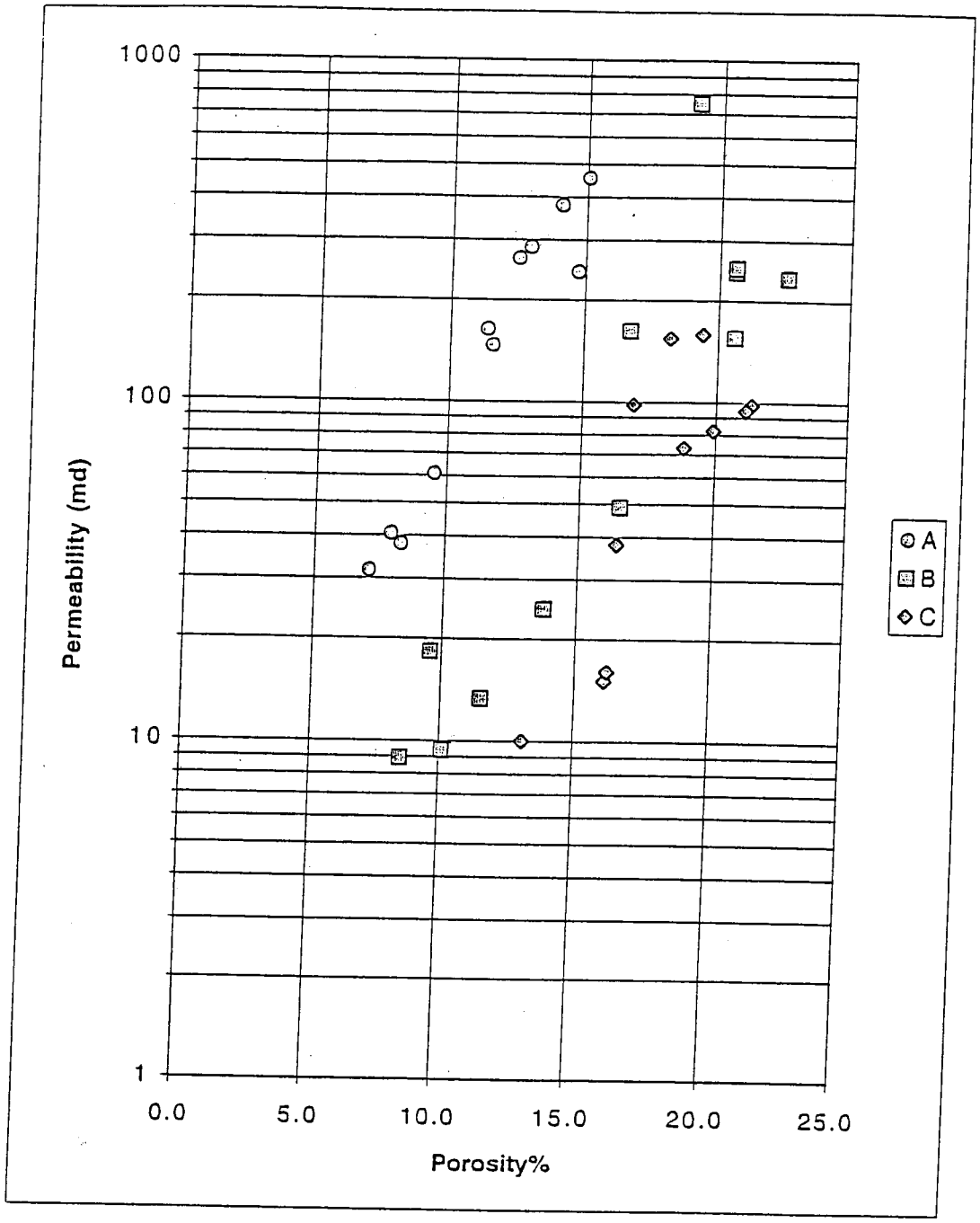


Figure 2. Crossplot of porosity and permeabilty of Simpson Sandstone core samples from Stafford County, Kansas measured by Alan Byrnes :medium-fine grained sandstone (petrofacies A), very fine-grained sandstone (petrofacies B), and very fine-grained/ silty sandstone (petrofacies C).

Petrofacies: porosity versus permeability
Simpson Sandstone sorted by grain size and basic lithofacies

ID#	% Porosity	Permeability md med-fn ss A	vfn ss B	/fn-slty ss C
18	15.0	455		
19	12.9	287		
20	14.0	379		
58	11.6	147		
59	14.7	243		
60	12.5	265		
61	11.4	164		
62	9.5	61.0		
63	7.1	31.7		
64	8.3	38.0		
65	7.9	40.7		
1	8.5		8.92	
2	11.5		13.3	
3	10.1		9.41	
4	13.8		24.5	
5	9.6		18.2	
43	20.6		155.0	
44	22.6		233	
45	20.6		243	
46	20.6		249	
47	16.5		48.8	
56	16.7		162	
57	19.1		751	
31	18.2			154
32	19.4			158
33	21.3			98.0
34	21.1			94.0
35	19.9			82.0
36	16.9			98.0
37	18.8			73.0
38	16.4			38.0
39	16.1			15.0
40	16.2			16.0
41	13.1			10.0

Petrofacies A mean values

Porosity	Permeability
11.4	132.9

Petrofacies B mean values

Porosity	Permeability
15.8	65.7

Petrofacies C mean values

Porosity	Permeability
17.9	54.3

Figure 3. Summary of porosity and permeabilty measurements of Simpson Sandstone core samples from Stafford County, Kansas (data from Alan Byrnes) :medium-fine grained sandstone (petrofacies A), very fine-grained sandstone (petrofacies B), and very fine-grained/ silty sandstone (petrofacies C).

2. Petrofacies sequence simulation

A Markov chain succession of petrofacies units can be generated as a Monte Carlo simulation, by the use of a random number generator in conjunction with a transition probability matrix. The general structure of the transition probability matrix takes the form of a probability table, where each row is matched with a petrofacies at some time increment and the cells contain the probability that the sequence will be in the petrofacies matched with the column at the following time increment. In this application, "time" is transformed to depth and an increment of 0.5 feet was chosen to match standard digital log sampling frequency. The probabilities in the matrix diagonal cells control the thickness distribution of each petrofacies.

In the workbook pfsim.xls, the transition probability matrix is under the control of the user. By changing the matrix input values, sequences with fining-upwards or coarsening-upwards trends can be generated and petrofacies modeled with varying thicknesses. The structure of the sequence generator is shown in Figure 4 (and pfsim.xls, worksheet 'markov'). The user enters values in the transitional probability matrix, with the condition that the probabilities in each row sum to unity, and an initial state of either A, B, or C. The sheet then generates a sequence of states as an ordered set from a an initial time of zero to 100 units, using an EXCEL random number generator of RAND() in conjunction with a look-up table generated from the transition probability matrix. The time series is then inverted to a depth sequence of petrofacies, where initial petrofacies is set at the Free-Water Level (FWL) of the reservoir and the successive incremental states are scaled at half-foot intervals above the FWL. The top of the succession in the workbook is at 50 feet above FWL. By this choice of range, the final model will scaled to represent transition zones and the lower parts of reservoirs that are at "irreducible" water saturation. The limitation in scale is appropriate for Kansas applications, but is also dictated by the poorer performance of Pittman equations in modeling saturations above the transition zone.

MARKOV CHAIN SEQUENCE SIMULATOR

Initial state (A, B, or C) = C

Transition probability matrix

P=

	A	B	C
A	0.8	0.2	0
B	0.1	0.7	0.2
C	0.1	0.3	0.6

Lookup table

	A	B	C	Z
A	0	0.8	1	1
B	0	0.1	0.8	1
C	0	0.1	0.4	1

t	X(t)	RAND	X(t+1)
0	C	0.071	A
1	A	0.499	A
2	A	0.167	A
3	A	0.072	A
4	A	0.795	A
5	A	0.738	A
6	A	0.068	A
7	A	0.135	A
8	A	0.221	A
9	A	0.274	A
10	A	0.798	A
11	A	0.943	B
12	B	0.781	B
13	B	0.556	B
14	B	0.165	B
15	B	0.686	B
16	B	0.051	A
17	A	0.6	A
18	A	0.074	A
19	A	0.435	A
20	A	0.904	B
21	B	0.57	B
22	B	0.614	B
23	B	0.583	B
24	B	0.507	B
25	B	0.973	C
26	C	0.46	C
27	C	0.262	B
28	B	0.658	B
29	B	0.769	B
30	B	0.5	B
31	B	0.438	B
32	B	0.789	B
33	B	0.286	B
34	B	0.637	B

0	0.1	0.4	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.8	1	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.4	1
0	0.1	0.4	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1
0	0.1	0.8	1

Height(ft)	
50	B
49.5	B
49	B
48.5	B
48	B
47.5	B
47	A
46.5	A
46	B
45.5	B
45	B
44.5	B
44	B
43.5	B
43	B
42.5	A
42	B
41.5	C
41	B
40.5	B
40	B
39.5	B
39	B
38.5	B
38	B
37.5	A
37	A
36.5	C
36	C
35.5	C
35	C
34.5	B
34	A
33.5	A
33	B

Figure 4. The structure of the transition probability matrix and Markov chain sequence generator on pfsim.xls, worksheet 'markov'.

3 .Allocation of pore volumes and 4. permeabilities to a simulated succession of petrofacies

At each depth increment of the simulation, the petrofacies is identified and a porosity assigned, using a random number to select a value based on the petrofacies mean value and standard deviation (summarized in Figure 5 and on. pfsim.xls, worksheet 'petrofacies'. By using the EXCEL function NORMSINV(RAND()), the random numbers conform closely to a normal distribution, which is used a the model to characterize the porosity distributions in the three petrofacies.

From a regression analysis of logarithmic permeability predicted from porosity (Figures 5 and 6; pfsim.xls, worksheet 'petrofacies'), parameters for the intercept (a) and the slope (b) of each petrofacies were applied in the prediction of a permeability for each increment in the simulation.

Petrofacies characterization
Mean and standard deviations of porosity
Regression analysis of log k = a+b*phi

ID#	Phi	k	log k
18	0.15042	454.514	2.657547
19	0.12932	286.979	2.457851
20	0.14041	378.64	2.578226
58	0.116	147	2.167317
59	0.147	243	2.385606
60	0.125	265	2.423246
61	0.114	164	2.214844
62	0.095	61	1.78533
63	0.071	31.7	1.501059
64	0.083	38	1.579784
65	0.079	40.7	1.609594
1	0.08479	8.91569	0.950155
2	0.11531	13.3326	1.124915
3	0.10092	9.41183	0.973674
4	0.1378	24.5016	1.389195
5	0.09561	18.2438	1.261115
43	0.206	155	2.190332
44	0.226	233	2.367356
45	0.206	243	2.385606
46	0.206	249	2.396199
47	0.165	48.8	1.68842
56	0.167	162	2.209515
57	0.191	751	2.87564
31	0.182	154	2.187521
32	0.194	158	2.198657
33	0.213	98	1.991226
34	0.211	94	1.973128
35	0.199	82	1.913814
36	0.169	98	1.991226
37	0.188	73	1.863323
38	0.164	38	1.579784
39	0.161	15	1.176091
40	0.162	16	1.20412
41	0.131	10	1

Petrofacies A	
Porosity	
mean	sd
0.114	0.028
a	b
0.4385	14.8274

Petrofacies B	
Porosity	
mean	sd
0.158	0.050
a	b
-0.0985	12.0928

Petrofacies C	
Porosity	
mean	sd
0.179	0.025
a	b
-0.7084	13.6126

Figure 5. Petrofacies statistics for porosity and permeability.

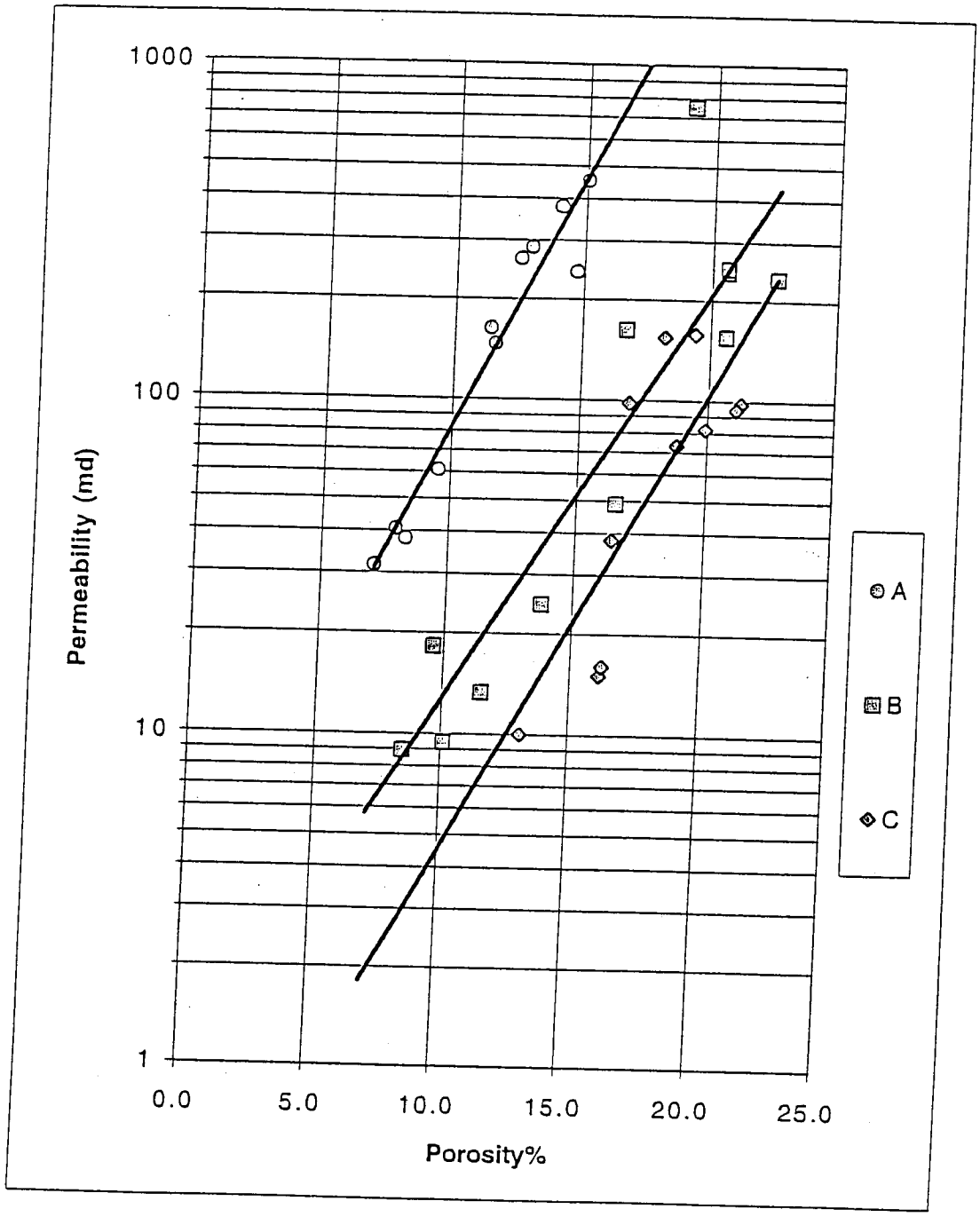


Figure 6. Permeability-porosity regression trends for petrofacies A, B, and C.

5. Pittman equation modeling of water saturations based on height above FWL, permeability, and porosity

Pittman (1992) made predictions of the radii of pore-throats penetrated over a range of mercury saturations from 10% to 75% in 5% increments. The measurements and statistical analysis were based on 202 sandstone samples from 14 formations, Ordovician to Tertiary in age, and with variation in composition and texture. Each equation was of the form:

$$\log(r_x) = A + B \cdot \log k + C \cdot \log \Phi$$

where r_X is the prediction of the radius of the smallest pore-throat penetrated when the sandstone is saturated by x% of mercury, k is the permeability of the sandstone in millidarcies, _ is the sandstone porosity in percent, and A, B, and C are constants determined by the regression analysis. The Pittman equation coefficients were used first to model capillary pressure curves based on the average porosity and permeability of the three petrofacies (Figure 7 and pfsim.xls, worksheet 'petrofacies'). The pore throat radius was converted to equivalent capillary pressure in a mercury/air system in units of psi, and then converted to equivalent height of oil column in feet, using a conversion number under user control.

Pittman (1992) noted that at the lower range of mercury saturation (10-35), the porosity term was not statistically significant and pore-throat radius could be predicted equally well by using permeability alone. Predictions in the range from 10 to 55% mercury saturation appeared to be the most reliable with correlation coefficients in excess of 0.90. Above 55% mercury, the regression fits declined progressively. Consequently the Pittman equations provide an adequate first-order model for sandstones, particularly in the transition zone, but predictions degrade at higher levels so that the simulation is best restricted to a limited interval above the FWL, and is fifty feet in pfsim.xls.

The Pittman equations are used in conjunction with the simulated succession by using the porosity and permeability at each level to compute a vector of expected heights above FWL that correspond to fixed increments of water saturation and then using the actual height above FWL to select the matching water saturation by an EXCEL look-up function as shown in Figure 8 (and pfsim.xls, worksheet 'markov').

Finally, the petrofacies assignments, porosities, and water saturations are combined together in a graphic profile of the simulated succession, ranging over a fifty-foot interval above the FWL (Figure 9 and pfsim.xls, worksheet 'markov').

PITTMAN'S EMPIRICAL EQUATIONS FOR SANDSTONES EXTENDED TO PREDICT HYDROCARBON COLUMN HEIGHT OVER A RANGE OF MERCURY SATURATIONS, BASED ON CORE POROSITY AND PERMEABILITY

	Porosity%	Permeability, md
A	11.4	132.9
B	15.8	65.7
C	17.9	54.3

PSI to hydrocarbon column conversion constant = 0.70

	Pore throat size (log r)				Height above FWL (feet)		
Hg%	A	B	C	Sw%	A	B	C
10	1.114	0.906	0.844	90	5.8	9.3	10.8
15	1.050	0.846	0.785	85	6.7	10.7	12.4
20	1.000	0.798	0.739	80	7.5	12.0	13.7
25	0.962	0.750	0.687	75	8.2	13.4	15.5
30	0.933	0.706	0.638	70	8.8	14.8	17.3
35	0.902	0.655	0.580	65	9.4	16.7	19.8
40	0.877	0.603	0.518	60	10.0	18.8	22.9
45	0.871	0.547	0.443	55	10.1	21.4	27.1
50	0.834	0.471	0.354	50	11.0	25.4	33.3
55	0.783	0.387	0.258	45	12.4	30.9	41.6
60	0.711	0.277	0.133	40	14.6	39.8	55.5
65	0.646	0.169	0.008	35	17.0	51.1	73.9
70	0.550	0.030	-0.147	30	21.2	70.3	105.8
75	0.398	-0.161	-0.353	25	30.1	109.1	170.0

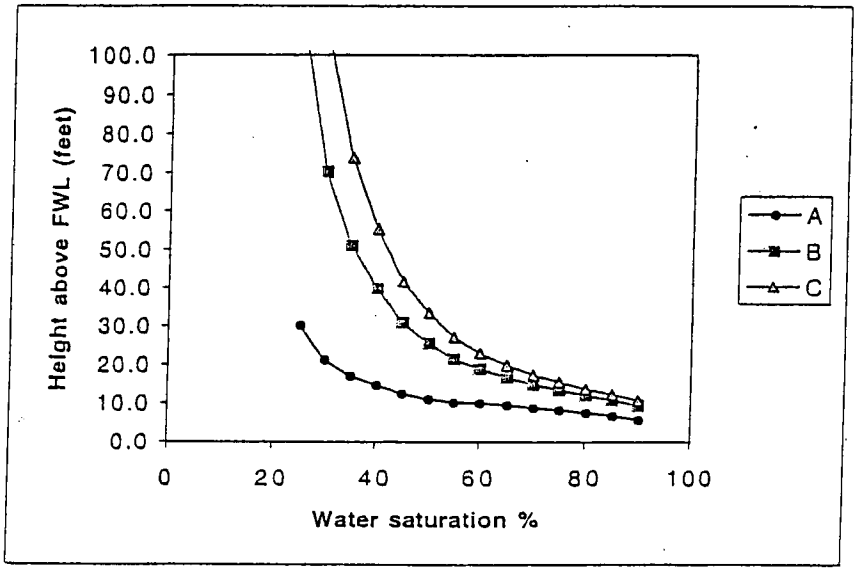


Figure 7. Use of the Pittman equation coefficients to model capillary pressure curves based on the average porosity and permeability of the three petrofacies.

PITTMAN EQUATIONS

Capillary pressure (psi) to height (feet) conversion = 0.7

Water saturation - height look-up table

Height	1	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5	0.45	0.4	0.35	0.3	0.25	SW
50	0	12.9	15.1	17.1	19.2	21.2	23.7	26.2	28.8	33.1	38.7	48.2	57.9	73.5	106	0.4
49.5	0	22.7	27.6	32.5	36.1	39.3	42.6	44.3	42.8	44.5	46.2	51.5	51.3	52.6	61.5	0.45
49	0	4.62	5.16	5.59	6.23	6.86	7.73	8.92	10.7	13.3	17.1	23.3	33.3	52.4	92.5	0.35
48.5	0	5.74	6.46	7.06	7.87	8.69	9.8	11.3	13.3	16.4	20.8	27.9	38.7	58.7	100	0.35
48	0	27.1	33.5	40.3	44.3	47.5	50.1	49.8	43.8	42.3	40.6	41.8	37	33.1	34.1	0.7
47.5	0	3.62	4.01	4.3	4.77	5.25	5.91	6.85	8.3	10.4	13.7	18.8	27.7	45.2	83	0.3
47	0	8.84	10.5	12.1	13.1	14	14.8	15.1	14.3	14.7	15.4	17.1	17.7	19.4	24.3	0.25
46.5	0	7.55	8.87	10.1	11	11.8	12.5	13	12.7	13.4	14.4	16.4	17.9	20.7	27.3	0.25
46	0	26.7	33	39.7	43.7	46.9	49.6	49.5	44	42.7	41.3	42.9	38.4	34.8	36.3	0.75
45.5	0	9.15	10.5	11.7	13.1	14.5	16.3	18.4	21	25.1	30.5	39.4	50.8	70.4	110	0.4
45	0	13.9	16.3	18.6	20.8	23	25.6	28.3	30.6	35	40.4	49.8	58.7	73.1	103	0.45
44.5	0	5.77	6.5	7.1	7.92	8.74	9.86	11.3	13.4	16.5	20.9	28	38.8	58.9	100	0.35
44	0	4.72	5.28	5.72	6.37	7.02	7.92	9.13	10.9	13.6	17.5	23.7	33.8	53	93.3	0.35
43.5	0	8.68	9.95	11.1	12.4	13.7	15.4	17.4	20	23.9	29.3	38.1	49.5	69.3	109	0.4
43	0	18.2	21.6	25.1	28	30.7	33.9	36.4	37.6	41.2	45.4	53.5	58.5	66.6	86.1	0.5
42.5	0	6.47	7.54	8.52	9.3	9.93	10.6	11.2	11.2	12	13.3	15.4	17.5	21.1	29.1	0.25
42	0	12.3	14.3	16.2	18.1	20.1	22.4	25	27.5	31.9	37.5	47	57.1	73.6	107	0.45
41.5	0	7.24	8.31	9.25	10.3	11.2	12.4	13.8	15.2	17.6	21.1	26.5	33.5	45.4	69.6	0.35
41	0	7.53	8.57	9.47	10.6	11.7	13.2	15	17.4	21.1	26.2	34.4	45.8	66.1	107	0.4
40.5	0	6.17	6.97	7.64	8.53	9.42	10.6	12.2	14.3	17.5	22.1	29.5	40.6	60.8	102	0.4
40	0	12.4	14.4	16.3	18.2	20.2	22.6	25.1	27.7	32	37.6	47.1	57.2	73.6	107	0.45
39.5	0	10.9	12.7	14.2	15.9	17.6	19.7	22.1	24.8	29.1	34.7	44.1	54.9	72.9	109	0.45
39	0	11.5	13.4	15.1	16.9	18.7	20.9	23.3	26	30.3	35.9	45.4	55.9	73.3	109	0.45
38.5	0	11.7	13.6	15.4	17.2	19	21.3	23.8	26.4	30.7	36.3	45.8	56.3	73.4	108	0.45
38	0	1.95	2.11	2.22	2.44	2.66	2.98	3.48	4.3	5.5	7.46	10.5	16.5	29.4	58.8	0.3
37.5	0	4.37	5	5.54	6.04	6.46	6.96	7.47	7.81	8.71	10.1	12.3	15	20	30.2	0.25
37	0	5.96	6.91	7.78	8.49	9.07	9.73	10.3	10.4	11.3	12.6	14.8	17.1	21.1	29.7	0.25
36.5	0	3.17	3.52	3.8	4.18	4.54	5.03	5.7	6.62	8.05	10.2	13.6	19.3	30.3	53.7	0.3
36	0	6.05	6.89	7.62	8.44	9.22	10.2	11.4	12.8	15	18.2	23.3	30.3	42.6	67.7	0.35
35.5	0	1.84	2.01	2.12	2.32	2.5	2.76	3.14	3.72	4.6	6.04	8.2	12.3	20.9	39.9	0.3
35	0	5.06	5.72	6.28	6.95	7.58	8.41	9.43	10.7	12.7	15.7	20.3	27.1	39.4	64.6	0.35
34.5	0	14.9	17.5	20.1	22.4	24.7	27.5	30.2	32.4	36.6	41.8	51.1	59.2	72.2	99.8	0.55
34	0	5.36	6.19	6.92	7.56	8.08	8.68	9.23	9.44	10.4	11.7	14	16.5	20.9	30.2	0.25
33.5	0	3.92	4.46	4.91	5.36	5.73	6.18	6.66	7.03	7.91	9.27	11.3	14.2	19.3	29.8	0.25
33	0	13.4	15.6	17.8	19.9	21.9	24.5	27.1	29.6	33.9	39.4	48.9	58.3	73.4	105	0.55
32.5	0	7.01	7.96	8.76	9.79	10.8	12.2	13.9	16.2	19.7	24.6	32.6	43.9	64.3	106	0.45

Figure 8. Use of the Pittman equations in conjunction with the simulated succession by inputting the porosity and permeability at each level in the computation of a vector of expected heights above FWL that correspond to fixed increments of water saturation and then using the actual height above FWL to select the matching water saturation by an EXCEL look-up function.

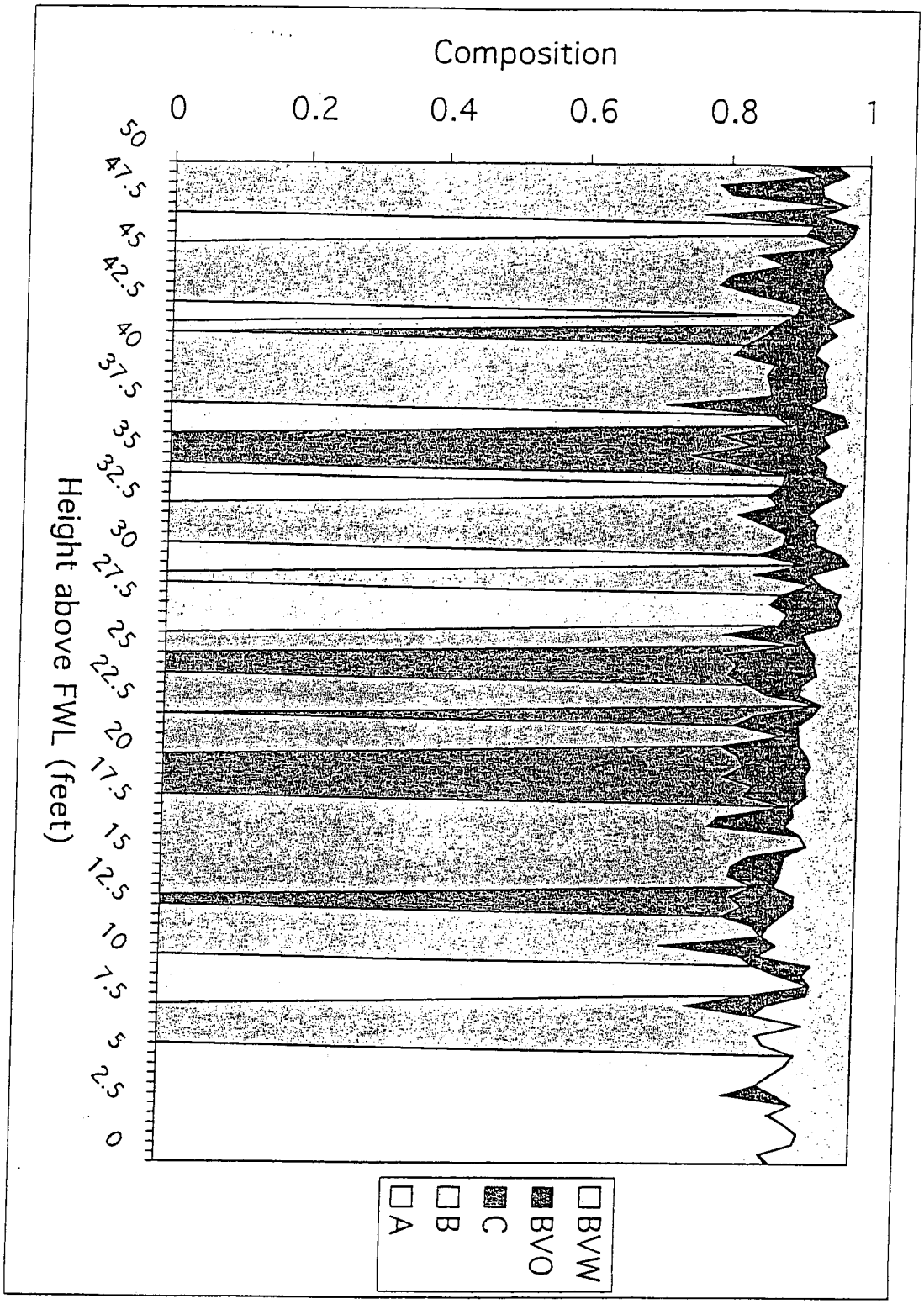


Figure 9. Compositional profile of simulation, graphed as petrofacies (A, B, or C), BVW (bulk-volume water), and BVO (bulk-volume oil).

6. Forward-modeling of resistivity curves and 7. generation of a Pickett plot

The Archie equation is applied to forward-model the resistivity log curve, using porosities and water saturations from the simulation. The values of the Archie equation parameters are selected by the user (Figure 10 and pfsim.xls, worksheet 'markov').

The porosity and resistivity values are then located on a Pickett plot (Figure 11 and pfsim.xls, worksheet 'markov') for evaluation in terms of the model and comparative work with real successions.

Conclusions

The software model presented in this report represents a first step in an integrated petrophysical evaluation, as a tool in forward-modeling of petrofacies to their expected log responses. In its current implementation, the workbook generates hypothetical successions, although they are controlled by the user's input of transition probabilities. However, future versions will be applied with real successions so that the workbook can be moved beyond a teaching tool to aid in the interpretation of Pickett plots from real log suites to predictive capabilities of FWL estimation and analysis of reservoir flow-unit architecture.

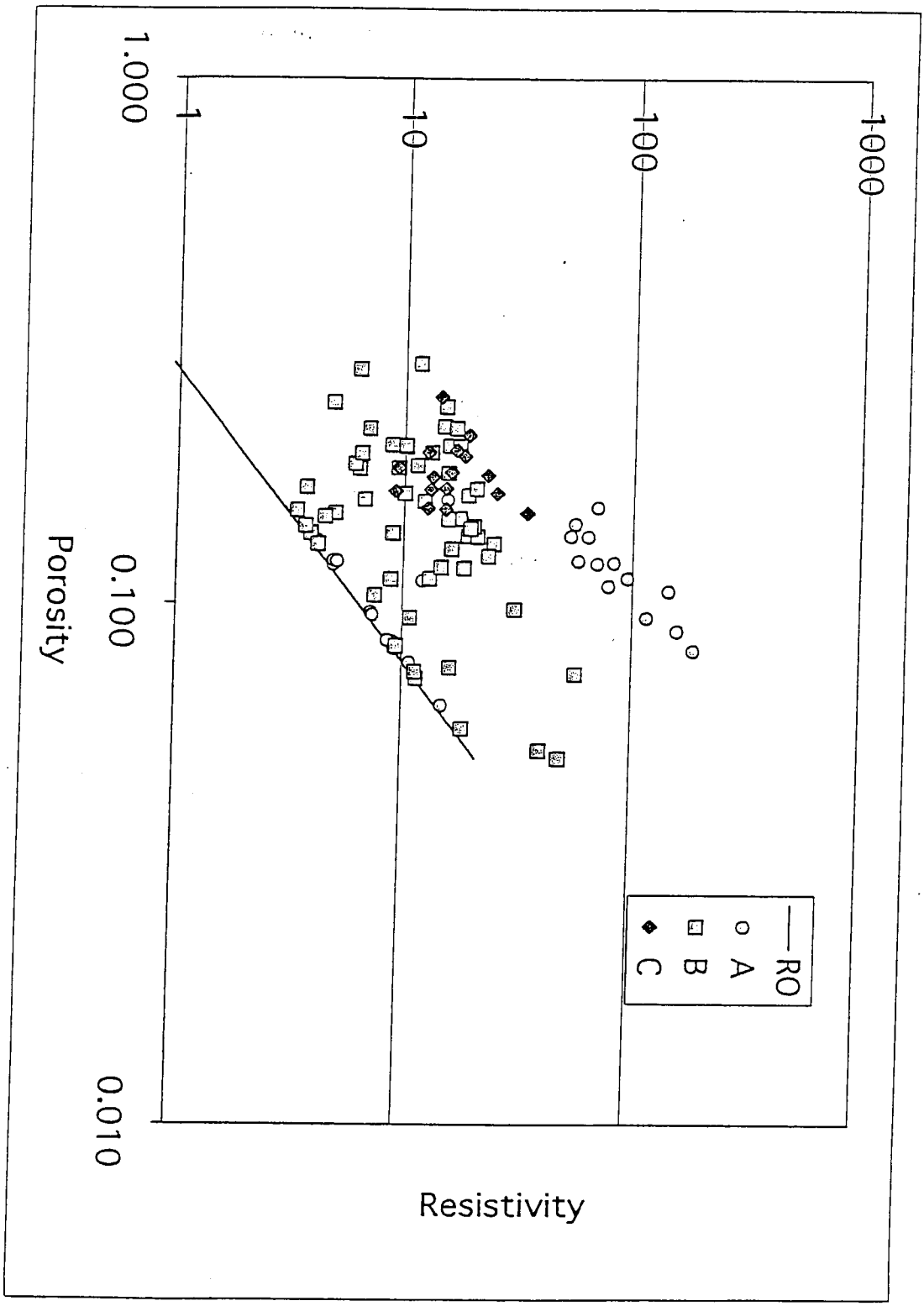


Figure 10. Pickett plot of resistivities and porosities from the simulation.

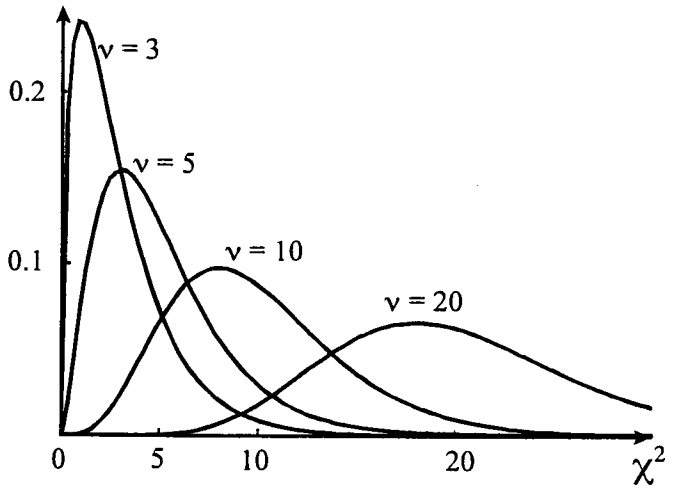
The chi-square distribution and chi-square test was introduced by Karl Pearson in 1900. It is related to the normal distribution where if n values are drawn independently from a normally distributed population with mean of μ and variance of σ^2 then:

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2$$

An important application of the chi-square distribution is in the chi-square test of goodness-of-fit that compares observed frequencies with frequencies that would be expected from a statistical distribution or from probability calculations. If O_i is the observed frequency in category i and E_i is the expected frequency then:

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i}$$

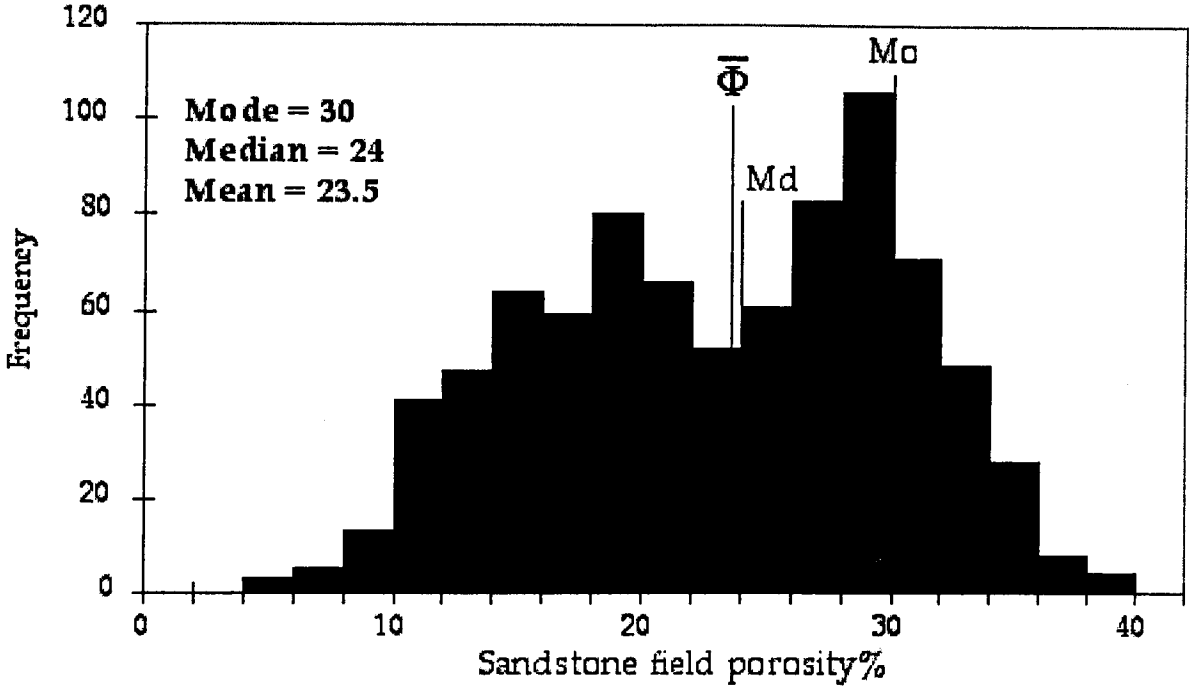
for k categories. The critical test value of chi-square is determined by the significance level chosen (conventionally $\alpha = 0.05$) and (k-1) degrees of freedom (ν). If the computed chi-square exceeds the critical value, then the null hypothesis of no difference between the observed and expected values is rejected. Critical values of the χ^2-distribution can be obtained using the CHIINV() function in EXCEL.



The χ^2 Distribution

EXAMPLE: PEARSON GOODNESS-OF-FIT χ^2 TEST

The distribution of porosities in the sandstone field of the TORIS database does not look normally distributed although it has a mean and a median that are very close. We can compute the frequencies that we would expect for these sandstone porosities, using the same mean and standard deviation.



TORIS DATABASE OF SANDSTONE RESERVOIR POROSITIES

COUNT	839
mean	23.4814462
stdev	7.31022862

PHI%	Z	PROP	EXPECTED	OBSERVED	(O-E)^2/E	INTERVAL
4	-2.6650	0.0038				
6	-2.3914	0.0084	3.8	3	0.2	1
8	-2.1178	0.0171	7.3	5	0.7	2
10	-1.8442	0.0326	13.0	13	0.0	3
12	-1.5706	0.0581	21.4	41	17.8	4
14	-1.2970	0.0973	32.9	47	6.1	5
16	-1.0234	0.1531	46.8	64	6.4	6
18	-0.7498	0.2267	61.8	59	0.1	7
20	-0.4762	0.3170	75.7	80	0.2	8
22	-0.2027	0.4197	86.2	66	4.7	9
24	0.0709	0.5283	91.1	52	16.8	10
26	0.3445	0.6348	89.4	61	9.0	11
28	0.6181	0.7317	81.4	83	0.0	12
30	0.8917	0.8137	68.8	106	20.1	13
32	1.1653	0.8780	54.0	71	5.4	14
34	1.4389	0.9249	39.3	48	1.9	15
36	1.7125	0.9566	26.6	28	0.1	16
38	1.9861	0.9765	16.7	8	4.5	17
40	2.2596	0.9881	9.7	4	3.4	18

First the class boundaries are converted to Z-scores by subtracting the mean porosity value and dividing by the standard deviation. Then the cumulative proportion of the fitted normal distribution at each boundary Z-score value is computed using the EXCEL function NORMSDIST(). The proportion in each interval that would be expected from the normal distribution is found by subtracting adjacent boundary cumulative proportions and, when multiplied by the total number of fields (839) gives the expected frequency in each interval.

When the chi-square value is calculated from:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

the value summed over the cells is 97.5. The number of intervals is 18 and so the number of degrees of freedom is 17 (n-1). The critical test value of chi-square at a significance level of 0.05 and 17 df is 27.6 . This is given by the EXCEL function CHIINV(0.05,17).

The computed figure exceed the critical test value and so the null hypothesis that the porosities of the sandstone fields in the TORIS database is normally distributed is rejected. The alternative, non-normal distribution looks distinctly bimodal, but systematic examination of this hypothesis is another issue.

EXAMPLE: PEARSON χ^2 TEST OF ASSOCIATION

In the TORIS database, there appears to be an association of the lithology of the fields (sandstone or limestone) and the API gravity of the oil.

TORIS DATABASE OF FIELD LITHOLOGIES AND API GRAVITIES

API	FIELD LITHOLOGY					
GRAVITY	SS	LS			SS	LS	
HEAVY	147	11	158		115.2	42.8	158
MEDIUM	165	35	200		145.8	54.2	200
LIGHT	406	221	627		457.0	170.0	627
	718	267	985		718	267	985

(O-E)^2/		
	8.8	23.7
	2.5	6.8
	5.7	15.3

CHI-SQUARE = 62.8

AT 2 DF AND ALPHA 0.05, CRITICAL CHI-SQUARE = 6.0

NULL HYPOTHESIS OF NO ASSOCIATION BETWEEN API GRAVITY AND RESERVOIR LITHOLOGY IS REJECTED

In this case, the expected frequencies are generated using the marginal probabilities in conjunction with the total number of fields to create a table that reflects no association between lithology and API gravity. The computation of chi-square from:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

gives a value of 62.8.

For a contingency table, the number of degrees of freedom is (r-1)*(c-1) where r is the number of rows and c is the number of columns. So, for this example, the number of degrees of freedom is 2. The critical test value of chi-square at a significance level of 0.05 and 2 df is 6.0 . This is given by the EXCEL function CHIINV(0.05,2).

The null hypothesis of no association between field lithology and API gravity is rejected in favor of the alternative hypothesis that there is an association. Remember that this says nothing about cause and effect between these two variables. In this instance, their association is probably related to a third variable, probably reservoir age, but this can be the subject of another statistical investigation.

REFERENCES

Ahmed, N., T. Natarajan, and H. R. Rainbolt, 1976, On generating Walsh spectrograms: IEEE Transaction on Electromagnetic Computability, v. EMC-18, no. 4, p. 198-200.

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

Alger, R.P., Raymer, L.L., Hoyle, W.R., and Tixier, M.P., 1963, Formation density log applications in liquid-filled holes: SPE Transactions of the AIME, v. 228, no. 3, p. 321-332.

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.

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

Beauchamp, K. G., 1984, Applications of Walsh and related functions: London, Academic Press, 308 p.

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

Chatfield, C., 1975, The analysis of time series: Theory and practice: London, Chapman and Hall, 263 p.

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

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.

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

Cooley, J. W., and J. W. Tukey, 1965, An algorithm for the machine computation of complex Fourier series: Mathematical Computation, v. 19, p. 297-301.

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.

Desbarats, A.,1989, Support effects and the spatial averaging of transport properties: Mathematical Geology, v. 21, no. 3, p. 383-389.

Doveton, J.H., 1971, An application of Markov chain analysis to the Ayrshire Coal Measures succession: Scottish Jour. Geol., v. 7, no. 1, p. 11-27.

Doveton, J. H., 1986, Log analysis of subsurface geology—concepts and computer methods: New York, John Wiley & Sons, 273 p.

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

Doveton, J. H., and Duff, P. McL. D., 1984, Passage-time characteristics of Pennsylvanian sequences in Illinois: Ninth ICC Congr. Compte Rendu, v. 3, p. 599-604.

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.

Gill, D., 1970, Application of a statistical zonation method to reservoir evaluation and digitized-log analysis: AAPG Bulletin, v. 54, no. 5, p. 719-729.

Gill, D., A. Shomrony, and H. Fligelman, 1993, Numerical zonation of log suites by adjacency-constrained multivariate clustering: AAPG Bulletin, in press.

Gingerich, P.D., 1969, Markov analysis of cyclic alluvial sediments: Jour. Sed. Petrology, v. 39, no. 1, p. 330-332.

Harmuth, H. F., 1977, Sequency theory: Foundations and applications: New York, Academic Press, 505 p.

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.

Jensen, J.L., Lake, L.W., Corbett, P.W.M., and Goggin, D.J., 2000, Statistics for Petroleum Engineers and Geoscientists, Elsevier, Amsterdam, 338 pp.

Kemeny, J.G., and Snell, J.L., 1960, Finite Markov Chains: Van Nostrand, Princeton, 210 pp.

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

Krumbein, W.C., and M.F. Dacey, 1969, Markov chains and embedded Markov chains in geology: Math. Geology, v. 1, no. 1, p. 79-96.

Lanning, E. N., and D. M. Johnson, 1983, Automated identification of rock boundaries: An application of the Walsh transform to geophysical well-log analysis: Geophysics, v. 48, no. 2, p. 197-205.

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

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

Matalas, N.C., 1967, Some distribution problems in time series simulation: Kansas Geol. Survey Computer Contr. 18, p. 37-40.

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.

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.

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.

Savre, W.C., 1963, Determination of a more accurate porosity and mineral composition in complex lithologies with the use of the sonic, neutron, and density surveys: Jornal of Petroleum Technology, v. 15, no. 6, p. 945-959.

Schwarzacher, W., 1967, Some experiments to simulate the Pennsylvanian rock sequence of Kansas: Kansas Geol. Survey Computer Contr. 18, p. 5-14.

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

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.

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.

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

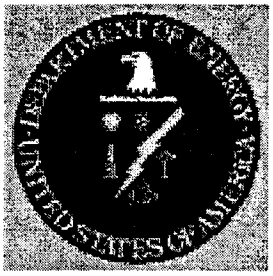
Weedon, G. P., 1989, The detection and illustration of regular sedimentary cycles using Walsh power spectra and filtering, with examples from the Lias of Switzerland: Journal of the Geological Society of London, v. 146, no. 1, p. 133-144.

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

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.

CPE940 Dataset 1

Tertiary Oil Recovery Information System (TORIS) data base



INTRODUCTION

NPC Public Database: (NPCPUBDB.GEO). Database developed by the National Petroleum Council (NPC) for its 1984 assessment of the nation's enhanced oil recovery (EOR) potential.
The technical data description is at the reservoir level. Included with the database are the Appendices from the "TORIS Data Preparation Guidelines" (NIPER/BDM-0042) defining the data elements in the database. Available in Spreadsheet format that can be used on a PC or Macintosh.

TORIS was originally developed by the National Petroleum Council (NPC) for its 1984 assessment of the nation's enhanced oil recovery (EOR) potential. The analysis was requested by the U.S. Secretary of Energy. In this effort, the EOR committee utilized and built upon data bases of individual oil reservoirs and computer models that were then under development by DOE, Office of Fossil Energy. After augmentation, adaptation, and validation, these oil reservoir data bases and models were remanded to the DOE's Bartlesville Project Office (BPO) for maintenance, updating, and subsequent application. The data bases and models become components of a larger system known as TORIS.

The TORIS data base currently contains over 2,540 oil reservoirs, accounting for over 64% of the original oil-in-place estimated to exist in discovered crude oil reservoirs in the U.S. TORIS utilizes its comprehensive data base and detailed engineering and economic methodologies at the reservoir level to estimate crude oil recovery, investment and operating costs, and ultimately project economics.

TORIS can analyze resource potential at two levels of technology: implemented and advanced. The implemented technology case assumes recovery processes that are currently available for implementation in the field. The advanced technology case assumes improvements in recovery technologies and reductions in extraction costs that will result from successful research and development (R&D) within a reasonable period of time. Each reservoir in the data base is subjected to a screening process to identify the technical applicability of alternative potential recovery processes.

TORIS is an analytical tool that has been utilized by DOE to support state agencies, federal agencies, Congress and industry by addressing broad policy issues in the areas of R&D, tax incentives, and environmental impacts. Through agreement with DOE, the Interstate Oil and Gas Compact Commission (IOGCC) has used the TORIS models and data base to evaluate the recovery potential by EOR and advanced secondary methods for its member States under a long-term project known as Advanced Oil Recovery and States.

TORIS.xls

TORIS.xls is a truncated subset of the complete TORIS database, where variables have been selected from the full set. However, all the fields (1269) have been retained.

Lithology Code. A numeric code describing the predominant lithology in the reservoir (Sandstone, Carbonate, or Dolomite). A distinction is made between a calcareous sandstone and a sandy limestone. The former is a sandstone and the latter is a limestone.

State Postal Code. A two-character postal abbreviation (e.g., CA, TX, etc.) of the name of the state in which the reservoir is located.

Field Name. The officially registered field name, with no abbreviation.

Reservoir Name and Formation Name. The Reservoir Name is not necessarily the same as the Formation Name. While reservoirs are almost always named after the geologic formation in which they reside, this is not the case in every situation.

Reservoir Acreage (Acres). The actual surface of the reservoir corrected for dipping, folding, faulting, or other distortions of the rock. Note that this definition may result in the reservoir acres being greater than the field acres for highly slanted or distorted formations. This is the acreage used in the volumetric calculation of OOIP.

Net Pay (Feet). That portion of the oil interval in the reservoir which is determined to have reservoir quality values of permeability and porosity. The methods of determining net pay cutoff limits and the means of measuring them are region specific and are generally based on prior experience. Do not include any gas zones in the net pay determination.

Gross Pay (Feet). The thickness of the entire oil interval in the reservoir including intervals which fall below the permeability and porosity standards used to determine net pay. This element is primarily used by the steamflood model in accounting for heat loss. Do not include any gas zones in the gross pay determination.

Porosity (%). Obtained from whole core studies or more commonly from electric log data. This should be the porosity value used in calculating Swi, from which Soi is obtained. The value must be a weighted average representative of the entire reservoir and must be greater then 7%. The source of the porosity data should be documented in the source file.

Initial Oil Gas and Water Saturation, (%). These values should be determined at reservoir conditions and should represent the entire reservoir. The values are usually derived from electric log analysis. The three saturations must sum to 100 percent.

True Vertical Depth (Feet). The distance from the Kelly Bushing to the mid-point of the perforations in the reservoir under consideration, expressed as a positive (not subsea) number. This should be a representative value for the entire reservoir.

Formation Temperature (ƒF). The best source for this datum is usually the maximum recorded temperature from the electric wireline logs or temperature logs.

Permeability (MD). The effective, dynamic, horizontal permeability of the reservoir in millidarcies. Preferentially this should come from whole core studies, but it may be calculated from pressure buildup test or sidewall core analysis.

API Gravity (ƒAPI). The initial producing API gravity of the oil, as specified in the American Petroleum Institute guidelines. This value should be taken from early producing data before the introduction of stimulating fluids which could alter the composition of the produced oil. It is permissible to approximate the gravity from field curves. This should be a representative value for the entire reservoir.

Formation Salinity (PPM TDS). The total dissolved solids in parts per million, best obtained from the downhole sampler.

OOIP (BBL). The OOIP for this system must be volumetrically derived. Since the TORIS models use rock and fluid properties to estimate tertiary recovery, the OOIP must represent those physical quantities. This means that the OOIP must be consistent with the volumetric data. OOIPs derived from material balance equations or from decline curves will not necessarily agree with the volumetric OOIP.

EXCEL TORIS.xls DATABASE
COLUMN VARIABLE

A State Postal Code (2 characters)
B Lithology Code (1=Sandstone; 2=Carbonate; 3=Dolomite) (2 digits)
C Geologic Age Code, AAPG (Three-digit integer code as shown in code sheet)
D Field Name
E Reservoir Name
F Net Pay (Feet)
G Gross Pay (Feet)
H Porosity (%)
I Initial Water Saturation (%)
J True Vertical Depth (Feet)---Mid-Perforation
K Formation Temperature (ƒF)
L Current Formation Pressure (PSI)
M Permeability (MD)
N API Gravity (ƒAPI)
O Formation Salinity (PPM TDS)
P OOIP (BBL)
Q Reservoir Acreage (Acres)
R Initial Formation Pressure (PSI)
S Depositional System (Three-digit integer code as shown in code sheet)

AGE CODES

-1	Unknown
100	Cenozoic
110	Quaternary
111	Holocene
112	Pleistocene
120	Tertiary
121	Pliocene
122	Miocene
123	Oligocene
124	Eocene
125	Paleocene
200	Mesozoic
210	Cretaceous
211	Cretaceous/Upper
212	Cretaceous/Gulf
213	Cretaceous/Coloradoan
217	Cretaceous/Lower
218	Cretaceous/Comanche
219	Cretaceous/Coahuila
220	Jurassic
221	Jurassic/Upper
224	Jurassic/Middle
227	Jurassic/Lower
230	Triassic
231	Triassic/Upper
234	Triassic/Middle
237	Triassic/Lower
300	Paleozoic
310	Permian
311	Permian/Upper
312	Permian/Ochoa
313	Permian/Guadalupe
317	Permian/Lower
318	Permian/Leonard
319	Permian/Wolfcamp
320	Pennsylvanian
321	Pennsylvanian/Upper
322	Pennsylvanian/Virgil
323	Pennsylvanian/Missouri
324	Pennsylvanian/Middle
325	Pennsylvanian/Des Moines
326	Pennsylvanian/Atoka
327	Pennsylvanian/Lower
328	Pennsylvanian/Morrow
330	Mississippian
331	Mississippian/Upper

332 Mississippian/Chester
333 Mississippian/Meramec
337 Mississippian/Lower
338 Mississippian/Osage
339 Mississippian/Kinderhook
340 Devonian
341 Devonian/Upper
342 Devonian/Chautauquan
343 Devonian/Senecan
344 Devonian/Middle
345 Devonian/Erian
347 Devonian/Lower
348 Devonian/Ulsterian
350 Silurian
351 Silurian/Upper
352 Silurian/Cayugan
354 Silurian/Middle
355 Silurian/Niagaran
357 Silurian/Lower
358 Silurian/Alexandrian
361 Ordovician/Upper
362 Ordovician/Cincinnatian
364 Ordovician/Middle
365 Ordovician/Champlanian
367 Ordovician/Lower
368 Ordovician/Canadian
370 Cambrian
371 Cambrian/Upper
372 Cambrian/Croixian
374 Cambrian/Middle
375 Cambrian/Albertan
377 Cambrian/Lower
378 Cambrian/Waucoban
400 Precambrian

Description of Geologic Reservoir Classification System

The following discussion will help familiarize the data preparer with the geological classification system used in TORIS. The classification in TORIS incorporates an individual assessment of the: (1) depositional system, (2) diagenetic overprint, and (3) structural compartmentalization, in order that the reservoir can be compared to other reservoirs with similar properties.

In practice, the primary decision in applying the classification first requires the determination of the lithology of the reservoir, i.e., carbonate or siliciclastic. Each lithologic type is secondarily characterized by the three basic elements as outlined in Figure 1. Each element axis includes a series of categories that are designed to include the range of most likely possibilities for that particular element but still be mutually exclusive. Each category has been further subdivided into sub-categories in order to capture more detailed facies information if it is available.

Definition and characteristics of individual categories of the element axes are based on current acceptable usage as defined in standard geologic texts. Boundary conditions between categories are gradational and by their very nature interpretive, thus creating a subjective element in the classification. However, the categories are made sufficiently broad in order to minimize differences in interpretation.

DEPOSITIONAL ENVIRONMENT CODES

-1 Unknown
0 Depositional System does not apply to heterogeneity
100 Default
110 Eolian
111 Eolian/Ergs
112 Eolian/Coastal Dunes
120 Lacustrine
121 Lacustrine/Basin Margin
122 Lacustrine/Basin Center
130 Fluvial Undifferentiated
131 Fluvial Braided Stream
132 Fluvial Meandering Stream
140 Alluvial Fan
141 Alluvial Fan/Humid
142 Alluvial Fan/Semi-Arid
143 Alluvial Fan/Fan Deltas
150 Delta/Undifferentiated
151 Delta/Wave-Dominated
152 Delta/Fluvial-Dominated
153 Delta/Tide-Dominated
160 Strandplain/Undifferentiated
161 Strandplain/Barrier Core
162 Strandplain/Barrier Shoreface

163 Strandplain/Back Barriers
164 Strandplain/Tidal Channels
165 Strandplain/Washover Fan/Tidal Delta
170 Shelf
171 Shelf/Sand Waves
172 Shelf/Sand Ridges/Bars
180 Slope-Basin (Clastic)
181 Slope-Basin/Turbidite Fans (Clastic)
182 Slope-Basin/Debris Fans (Clastic)
190 Basin (Clastic)
191 Basin/Pelagic
220 Peritidal
221 Peritidal/Supratidal
222 Peritidal/Intertidal
223 Peritidal/Subtidal
230 Shallow Shelf
231 Shallow Shelf/Open
232 Shallow Shelf/Restricted
240 Shelf Margin
241 Shelf Margin/Rimmed Shelf
242 Shelf Margin/Ramps
250 Reefs
251 Reefs/Pinnacle
252 Reefs/Bioherms
253 Reefs/Atolls
260 Slope-Basin (Carbonate)
261 Slope-Basin/Debris Fans (Carbonate)
262 Slope-Basin/Turbidite Fans (Carbonate)
263 Slope-Basin/Mounds
270 Basin (Carbonate)
271 Basin/Drowned Shelf
272 Basin/Deep Basin

CPE940 Dataset 2

Picaroon Sandstone reservoir petrography/petrophysics database

PICAROON SANDSTONE

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.

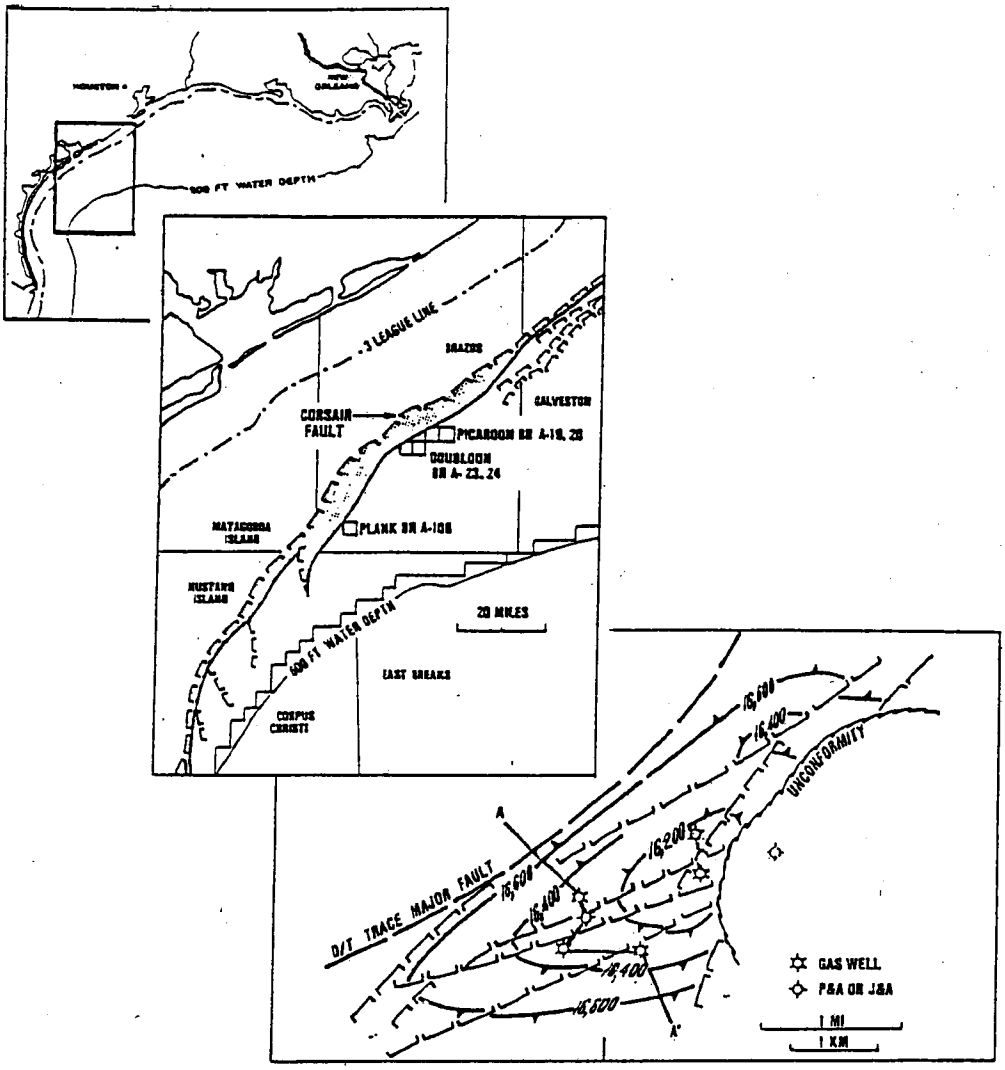
Taylor (1990) reported thin-section point counts of composition together with porosity and permeability for Picaroon field sandstones. The compositional variables recorded were: quartz, lithic fragments, detrital calcite (limestone rock and minor marine shell fragments), clay, quartz cement, calcite 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.

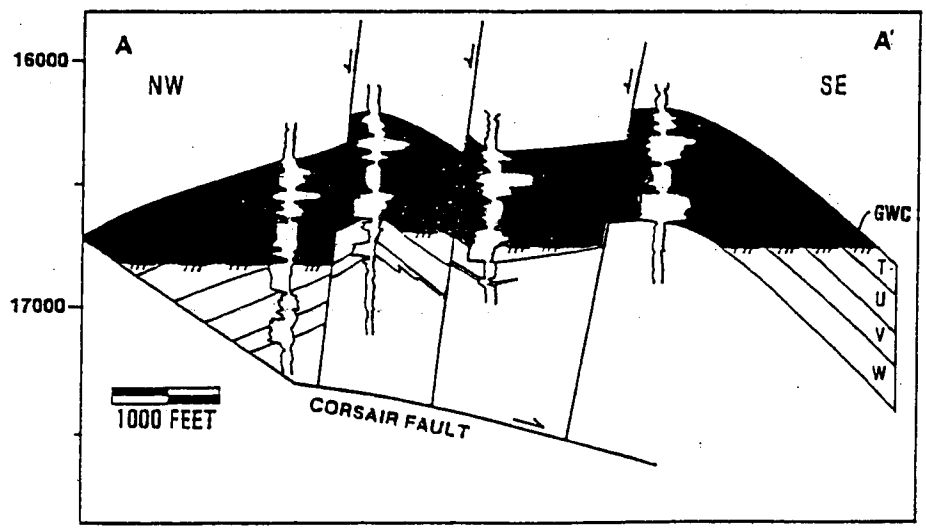
The depth range of the 39 sandstones show distinctive groupings at four levels. Four informal subgroups have been named as "Unit 0" (Sample #1), "Unit 1" (samples #2 to #21), "Unit 2" (samples #22 to 29), and "Unit 3" (samples #30 to #39). In the following demonstrations of statistical techniques, the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

REFERENCES

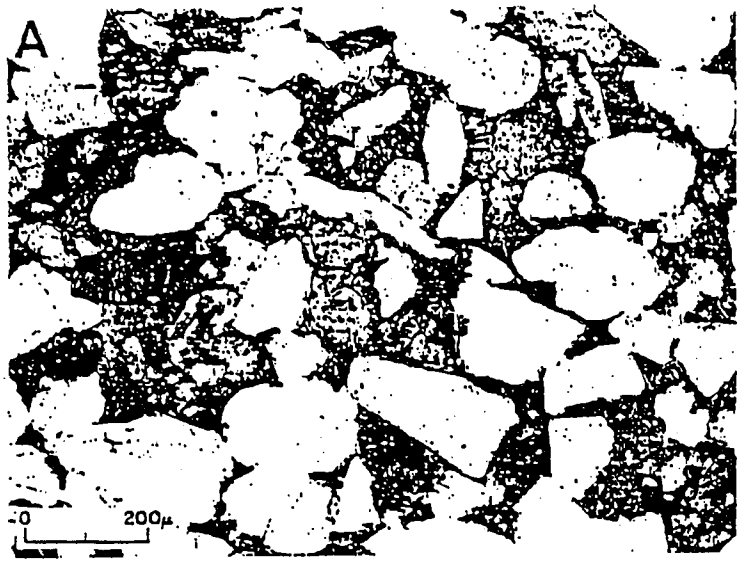
Taylor, T.R., 1990, The influence of calcite dissolution on reservoir porosity in Miocene sandstones, Picaroon field, offshore Texas Gulf Coast: Jour. Sed. Petr., v. 60, no. 3, p. 322-334.

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)



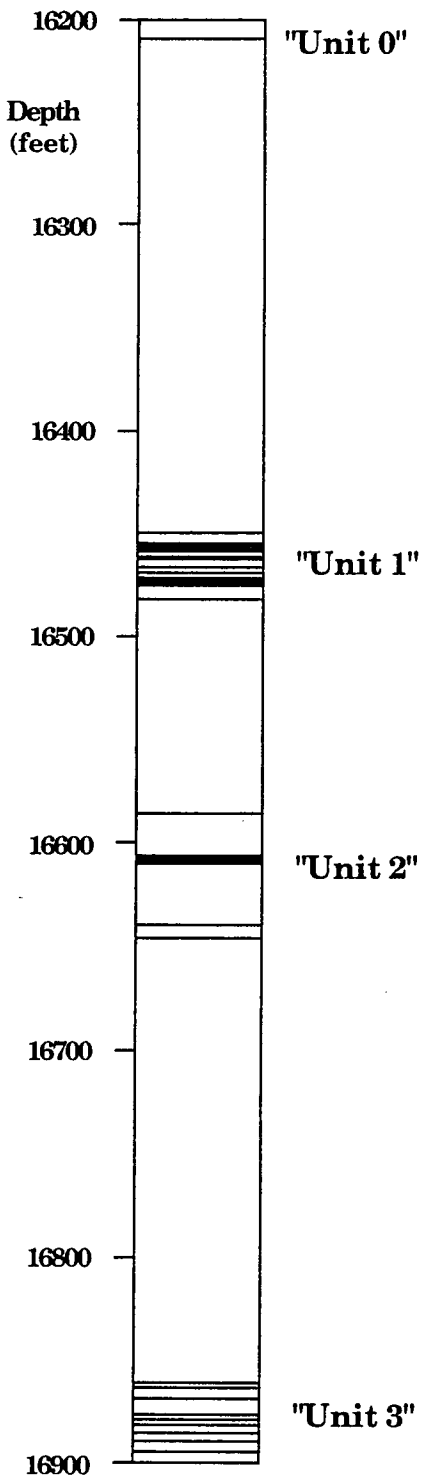
A. A low-porosity sandstone with approximately 30% calcite cement.

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 Field sandstone thin-sections (from Taylor, 1990)

PICAROON SANDSTONE (PICAROON GAS FIELD, MIOCENE, OFFSHORE TEXAS)							
ID#	DEPTH	PHI%	PERM md	QUARTZ%	LITH%	CLAY%	CEMENT%
1	16209.0	19.6	43	39.2	19.3	4.5	17.4
2	16450.0	23.3	525	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	4.9	41.1	23.1	0.6	17.0
5	16455.5	17.6	2.6	34.4	23.7	1.0	23.3
6	16456.0	19.6	2.9	46.3	15.5	1.5	17.0
7	16457.0	19.8	3.8	44.9	20.2	0.3	14.8
8	16457.5	21.2	5.4	36.2	26.6	1.6	14.4
9	16458.0	22	87	40.8	25.5	0.3	11.4
10	16459.0	21.4	73	40.2	22.2	2.4	13.8
11	16459.0	21.4	73	38.5	25.7	0.6	13.8
12	16461.0	18.2	1.8	42.0	25.4	3.6	10.7
13	16462.0	20	5.3	40.2	25.9	3.8	10.1
14	16463.0	20.4	6.9	44.4	19.8	2.6	12.8
15	16466.5	12.5	4.6	36.5	21.4	6.2	23.4
16	16469.0	15	21	34.1	22.2	6.4	22.2
17	16472.0	16.8	7.8	33.7	25.3	1.7	22.5
18	16473.0	17.2	7.2	35.8	22.8	3.4	20.8
19	16474.0	19.3	17	31.9	25.3	2.5	21.0
20	16475.0	11.1	1	37.4	25.1	2.0	24.4
21	16482.0	14.1	5.3	34.4	21.8	1.6	28.0
22	16586.0	10	0.01	38.6	19.1	5.8	26.5
23	16606.0	19.5	71	38.0	20.9	4.2	17.3
24	16607.0	20	82	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	2.4	37.7	28.9	3.3	16.4
27	16610.5	9	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	0.03	47.1	22.1	0.6	28.2
32	16868.4	11.2	2.1	51.6	20.6	0.3	16.2
33	16869.0	21.1	215	50.4	12.2	3.3	12.9
34	16876.0	28.2	896	43.8	22.2	1.6	4.1
35	16879.0	27.8	1150	50.2	15.5	2.5	4.1
36	16881.0	27.4	1210	48.1	19.0	1.3	4.2
37	16886.0	27.2	1100	49.8	17.2	2.4	3.4
38	16890.0	25.7	399	44.6	23.5	1.9	4.3
39	16895.0	27.1	838	49.9	18.1	1.6	3.3

Picaroon Sandstone data table



Picaroon Sandstone sample depth locations