OIL AND GAS EXPLORATION: STATISTICAL DECISION CRITERIA

THOMAS K. WIGNALL

C.1

OIL AND GAS EXPLORATION:

STATISTICAL DECISION CRITERIA.

BY

THOMAS KENNETH WIGNALL.

A THESIS SUBMITTED TO THE COMMITTEE ON GRADUATE STUDIES

IN PARTIAL FULLFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS.

DEPARTMENT OF MATHEMATICS

MEMORIAL UNIVERSITY OF NEWFOUNDLAND

ST. JOHN'S, NEWFOUNDLAND.

AUGUST, 1967.

ACKNOWLEDGEMENTS

I am indebted to Dr. Jean Tague of the Mathematics Department of Memorial University of Newfoundland who directed my line of research in statistical decision functions and discriminatory analysis.

Dr. Richard Hughes and Dr. W.G. Smitheringale of the Geology Department gave me valuable assistance with the petroleum geology for which I am most grateful. Dr. Hughes also used his wide experience in the petroleum exploration field to help with the interpretation of results.

I should like to thank Messrs. Sproule and Associates of Calgary for supplying valuable field data.

T.K. Wignall.

MEMORIAL UNIVERSITY OF NEWFOUNDLAND MATHEMATICS DEPARTMENT.

OIL AND GAS EXPLORATION: STATISTICAL DECISION CRITERIA: SUMMARY

by Thomas Kenneth Wignall.

Bayesian a priori probabilities are applied in the field of petroleum exploration to give the optimum decision criteria in locating oilwells and oil-fields.

Principal Component functions and discriminant functions are defined which make use of information available: geological, geophysical, or geochemical. The field studies confirm that these functions are very valuable in discriminating between producers and non-producers, achieving up to 95% success as the results given in the appendix prove.

The principal component scores and discriminant scores may be allotted to control points (oil and gas wells) on a map. Contours may then be mapped using the figures as probability indices. Thus new wells, fields, basins and provinces might be discovered, since these maps could be used along with structural contour maps to pinpoint new wells with a high probability of success.

The following functions defined in the thesis are all new:
(1) A favourability factor, F, using saturation ration, x_3, and shaliness, x_2, where $F = (x_3-1.5)(3.0-x_2)$, should prove most useful in helping to discover stratigraphic and hydrodynamic traps; also in deciding whether to complete a well.
(2) Principal Component Functions which diagnose what weight should be given to each variate responsible for the deposits of petroleum. This function is similar to the one given by Krumbein but is more powerful. A map using Principal components scores should help in the discovery of new resources.

(3) Discriminant functions are defined which are up to 95% effective in discrimination between dry holes and producing wells. Discriminant scores provide the most useful mapping. The field studies indicate that the data of petroleum wells is particularly amenable to discriminatory analysis; also the key variate or variates become very apparent, when an appropriate test is carried out.

Conclusion: A field study should now be carried out using the criteria defined. Information is difficult to collect as the Petroleum companies quite obviously do not wish to divulge any data which would aid their competitors. However, any data supplied to me will be treated as strictly confidential; and I will process the data and supply results and conclusions to any interested bodies who are willing to participate in the project. The more control points (wells) we have, the more useful the results will be. The data I require are two sets of stratigraphic or geophysical statistics from each field or basin: a set of producing wells and a set of non-producers. This is the project which I am now working upon, as a follow-up to this thesis.

CONTENTS

ERRATA:

Oil and Gas Exploration: Statistical Decision Criteria, by

Thomas Kenneth Wignall.

1. Page 33, paragraph 2, line 1, and paragraph 3, line 8:
for $BSB'/(n_1+ n_2 -2-p+1)$ read $BSB'(n_1+n_2-2)/(n_1+n_2-2-p+1)$.

2. Page 33, paragraph 3, Theorem 3.1, lines 1 and 9 :
for $(\overline{Y}_1 - \overline{Y}_2)/(n_1+n_2-2-p+1)$, read $(\overline{Y}_1-\overline{Y}_2)(n_1+n_2-2)/(n_1+n_2-2-p+1)$.

3. Page 44, paragraph 2, line 2:
for $(\overline{X}_1 - \overline{X}_2)/(n_1+n_2-2-p+1)$, read $(\overline{X}_1-\overline{X}_2)(n_1+n_2-2)/(n_1+n_2-2-p+1)$.

4. Page 44, paragraph 2, lines 5,6:
for 1.698/(22-2-3+1) read 1.698(21)/(22-2-3+1)= 1.98. Thus the standard error for individual wells is = 1.41.

5. Page 44, paragraph 3, line 2: for .849/.31=2.74 read .849/1.41=.602.

6. Page 44, paragraph 3, line 4 : for .0031, read .2735.

7. Page 44, paragraph 3, line 6 : for 99.7% read 72.65%.

8. Page 48, paragraph 2, lines 2,3 :
for $\sqrt{(6.34 - 4.07)/16} = .377$, read $\sqrt{(6.34-4.07)(19)/16} = 1.68$,
and for $\frac{1}{2}(6.34 -4.07)/.377$, read $\frac{1}{2}(6.34 - 4.07)/1.68 = .677$.

9. Page 48, paragraph 2, line 5 : for 99.8% read 74.9%.

T K Wignall

T.K.WIGNALL.

INTRODUCTION

This research arose from suggestions in Dr. Kaufman's work: "Statistical Decision and Related Techniques in Oil and Gas Exploration". Dr. Kaufman posed the following research problems on which my work has been concentrated:"(1) The use of polynomial regression schemes to isolate regional from local effects has recently come into vogue among geologists and geophysicists.Thus far the technique has been regarded as a technical tool to enable the expert to understand the geological nature of an area where well control is fairly good. However the map - be it gravity, magnetic, seismic,isopach, or facies - is also one of the chief types of information the operator has available to him in economic decision making; the real purpose motivating understanding of the purely technical (geological) aspects of an area. In a majority of instances the yield in barrels of oil or MCF of gas is directly proportional to the thickness of the sand stratum containing the hydrocarbon deposits.Sand thickness is portrayed on an isopach map. Clearly a regression scheme can be used to predict the expected values of sand thicknesses that may be encountered in a borehole within the limits of the area analysed. One might also wish to know whether further experimentation is warranted before proceeding with a test drill hole. (2) The devices used to gather information in reconnaissance exploration are highly imperfect in two senses. First, there is the possibility that the device will yield misleading information: e.g. a seismograph may indicate a structural high where there is none. Secondly, geological and geophysical tests give information pertaining to conditions favourable to the accumulation of oil and gas; they never tell whether oil or gas is present. Only the drill can confirm the existence of oil or gas; and devices are expensive; eg. seismic crew costs up to $5,000. per day. Can information theory assist the operator in coming to a decision?"

In this thesis we shall examine ways of discussing these and other vital questions with which a petroleum operator is faced.We shall use Bayesian techniques to give decision criteria which maximise the expected utility (profit). We shall formulate optimum decision discriminatory functions based on the information available which will give a discriminant score to each control point, and at the same time will yield partial scores such as isopach, and favorability factor scores, thus several useful mappings may result from one discriminant function. This plan of using a set of producers and a set of non-producers should provide stronger contours for deciding where to drill than any previously defined. We shall also use component analysis to formulate the function which determines the most suitable measures in predicting the presence of commercial quantities of petroleum.

In "Geology of Petroleum", the late Dr. Levorsen noted "the essential elements of oil and gas are simple: a porous, permeable rock called the reservoir rock, is overlain by an impervious rock, called the roof rock, contains oil or gas, the rock being deformed or so constructed that the petroleum is trapped.The simplest classification of petroleum deposits is based on mode of occurrence: surface or subsurface.Many of the major petroleum provinces were discovered as a result of a surface showing of oil, gas, or asphalt, since these occurrences indicate the presence of a source rock. Petroleum seepages are common in the sedimentary regions of the world , and they are frequently associated with water springs. Gas seepages are more readily observed when they occur in swamps or through water". This then is therefore a direct method of searching for oil and gas. Is Newfoundland a potential petroleum province? Two holes were drilled in the Port-au-Port peninsula in 1965 and reached a total depth of 4917 feet in the Ordovician.The Grand banks region is now being explored, with several holes reaching a depth of 5000 feet. This however is a true wildcat region and being below the sea the chance of success is only 1 in 20 for each hole. Thus the chance of the first three being dry is $(0.95)^3$, approximately 0.854; thus the chance that one at least will produce is about 0.146, about one chance in 7. Nevertheless the information these test-bores give could prove of economic importance. For even if they are commercially unsuccessful, they may indicate the presence of a source-rock, and Newfoundland would then emerge as a petroleum province. To summarise, oil and gas exploration is normally only carried out in provinces where source rock is known to exist. Throughout this thesis we shall assume that this is the case, and that the probability of finding oil or gas in commercial quantities in a wildcat zone is 0.1 on land and 0.05 under water. We shall also limit our study to regions where there has been no proximate surface show of hydrocarbons since this would bias considerably the probability of success. The combination of a positive seismic anomaly and the known presence of petroleum indicate a much higher chance of discovering a commercial quantity of petroleum. We assume then the operator is drilling in petroliferous territory, and we examine the optimal decisions he should make.

DRILLING DECISION CRITERIA

The precise locating of a petroleum well is really a problem in applied probability.Using all available information we shall employ Bayesian techniques of utilising 'a priori' probabilities to maximise the expected utility (average return per well). A wildcat petroleum explorationalist must decide whether to drill an exploratory hole on a site or whether to sell his drilling rights on the land. He might also have to consider returning some land to the Provincial government under the "checker-board" regulations. He may also decide to carry out a geological and/or geophysical programme of investigation such as seismic recordings. If obtained they should give accurate information as to the geological structure of the underlying strata.

In order to arrive at a decision, we shall calculate the expected return that results from every decision. We shall then make the decision which maximises our expected return. To make the procedure perfectly clear we shall first give examples before proceeding to the general construction..

EXAMPLE 1.1. Decision: (i) Drill exploratory hole.

(ii) Sell drilling rights, do not drill.

State of world (underlying subsurface hydrocarbon contents):

(i) Oil in commercial quantity.

(ii) Dry hole.

Note: "Oil in commercial quantity" is a variable quantity which is distributed log-normally (Kaufman); it also depends upon the depth of drilling and the cost of producing.Thus the state of the world is a continuous variate Θ, and $U(\text{drill}) = \int u(\text{drill}|\Theta)\, p(\Theta)\, d\Theta$.

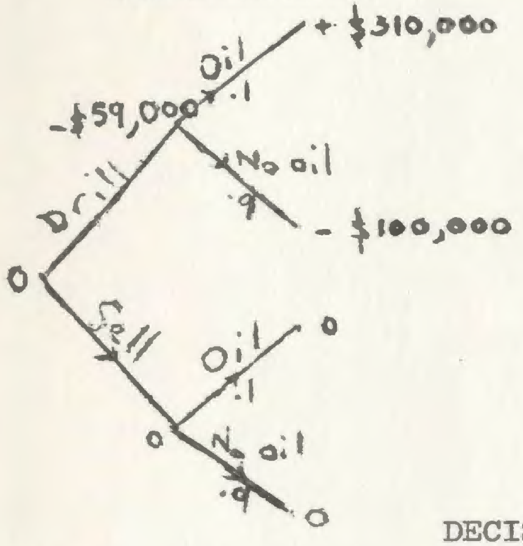
The average price of drilling a well is \$100,000,

Seismic and other information costs \$10,000,

The average producing well yields a profit of \$310,000.

Method:We assume a priori probabilities are 0.1 (producers),0.9 (not).

Decision Tree.



Calculations of expected returns resulting from each decision:

Decision (i) Drill. Expected profit=

(.1)(+\$310,000) + (.9)(-\$100,000) =

-\$59,000.

Decision (ii) Sell drilling rights.

Expected profit = 0.

Thus the maximum expected return = 0.

DECISION: SELL DRILLING RIGHTS, DO NOT DRILL.

Throughout the remainder of the chapter,we shall consider the following spaces,their elements, and a priori probabilities:

<u>Space</u>	<u>Elements</u>	<u>Interpretation</u>
D (1st Decision)	d_o	Do not take seismic readings
	d_1	Take seismic readings.
O (outcome of d_1)	o_o	No structure.
	o_1	Open structure.
	o_2	Closed structure.

Space	Elements	Interpretation
A (action: 2nd decision)	a_o	Do not drill,sell.
	a_1	Drill, retain 50% of shares.
	a_2	Drill, retain 100% of shares.
S (state of underlying	s_o	No oil.
structure)	s_1	Oil in commercial quantity.

Probabilities in a Wildcat Zone

$P(o_o) = .5$	$P(s_o\|o_o) = .9$	$P(s_1\|o_o) = .1$
$P(o_1) = .3$	$P(s_o\|o_1) = .7$	$P(s_1\|o_1) = .3$
$P(o_2) = .2$	$P(s_o\|o_2) = .4$	$P(s_1\|o_2) = .6$

In example 1.2 we shall use the decision tree method to maximise the expected utility when the operator has a set of two decisions to make.

EXAMPLE 1.2. D (i) d_o O (i) o_o A (i) a_o S (i) s_o
(ii) d_1 (ii) o_1 (ii) a_2 (ii) s_1
(iii) o_2

Method: In example 1.1, we examined the best decision to take when drilling was carried out without seismic information, and we saw that the maximum expected utility was 0. We will now investigate the expected utility when the decision is :undertake seismic survey.

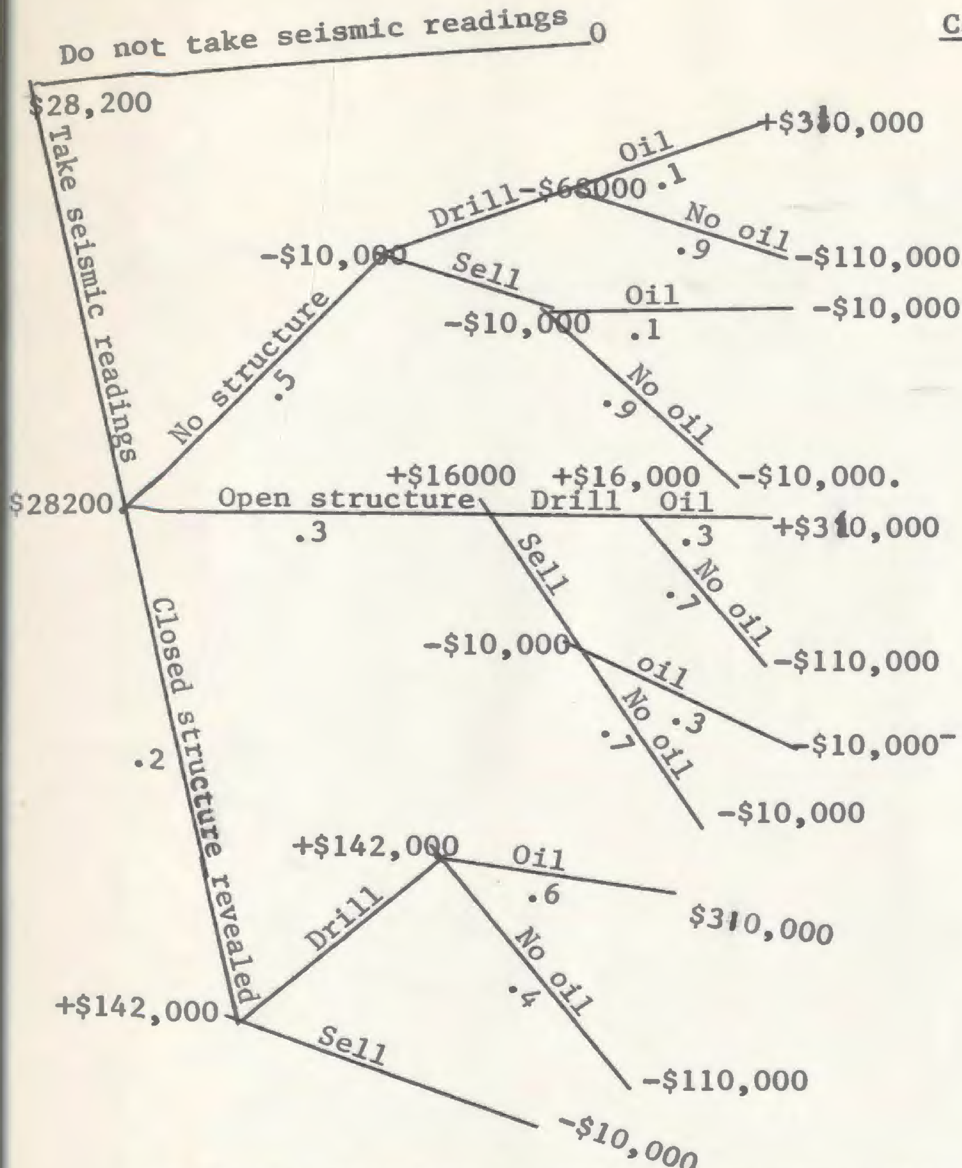
Decision Tree.

See P. 8

Calculations of expected returns.

"No structure" Expected profit from drilling = (.1)(+\$310,000) + (.9)(-\$110,000) = -\$68,000.

and from selling = -\$10,000.



__Calculations.__

Open structure:Expected profit from drilling = (.3)($310,000) + (.7)(-110,000) = __+$16,000__

and from selling= __-$10,000__.

Closed structure:Expected profit from drilling = (.6)(310,000) + (.4)(-$110,000) = __+$142,000__

and from selling= __-$10,000__.

Expected profit = (.2)(+142,000)+(.3)(+16,000)+(.5)(-$10,000) = __+$28,200__.

__Decisions__

Take seismic readings, and drill if open or closed structure is is revealed, otherwise sell the location drilling rights. Maximum Expected Utility = __+$28,200.__

In actual practice, petroleum explorationists often share out the risk elling shares and retaining a part interest in each hole drilled, since it ires a large capital to drill each well; and with a probability of success only there is a $(.9)^5$ chance of drilling 5 dry holes in succession, so that the ce of one or more producers is only .40951; hence one must spread the risk. In ple 1.3, we illustrate the procedure using all the spaces, elements and 'a priori' abilities defined on pages 6 and 7.

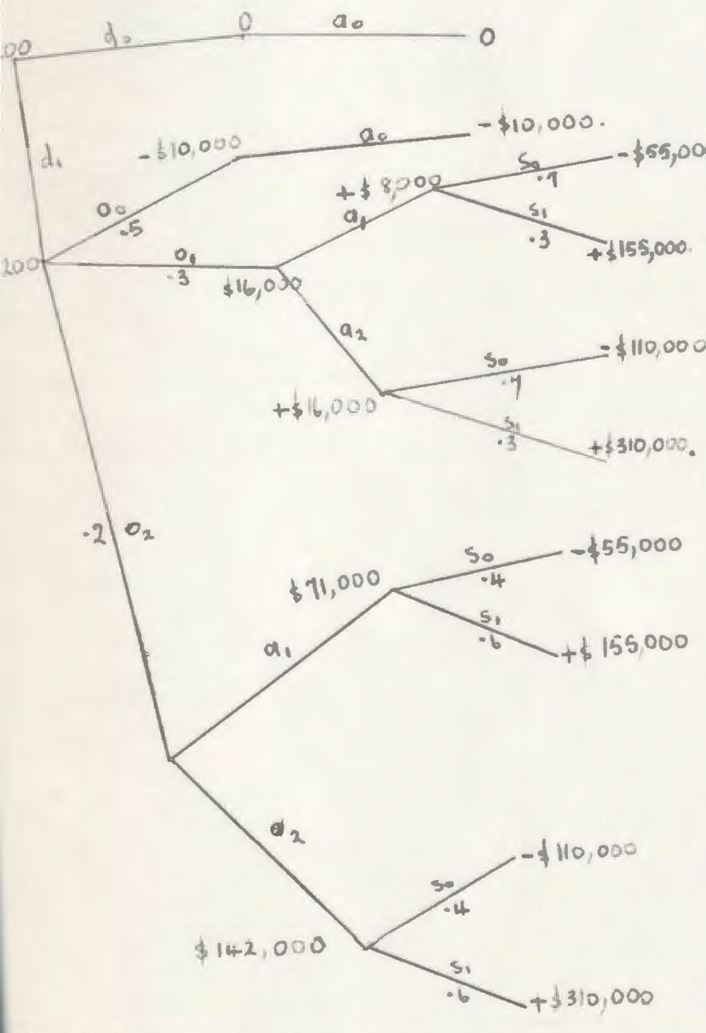
PLE 1.3. D (i) d_0 O (i) o_0 A (i) a_o S (i) s_o

(ii) d_1 (ii) o_1 (ii) a_1 (ii) s_1

(iii) o_2 (iii) a_2.

Method: The first decision is clear; take seismic readings. Also if the outcome is 'no structure', we sell drilling rights as in example 1.2. Hence we need to investigate the decisions required to maximise the expected return, if the outcome of the first decision is 'open structure' or 'closed structure'.

Decision Tree.



Calculations of expected returns.

o_1, a_1: Expected profit =(.5)(.3)(+$310,-000) + (.5)(.7)(-$110,000) = +$8,000.

o_1,a_2 Expected profit = (.3)(+$310,000) + (.7)(-$110,000) =+$16,000.

o_2,a_1: Expected profit = (.5)(.6)(+310,000) + (.5)(.4)(-110,-000) = +$71,000.

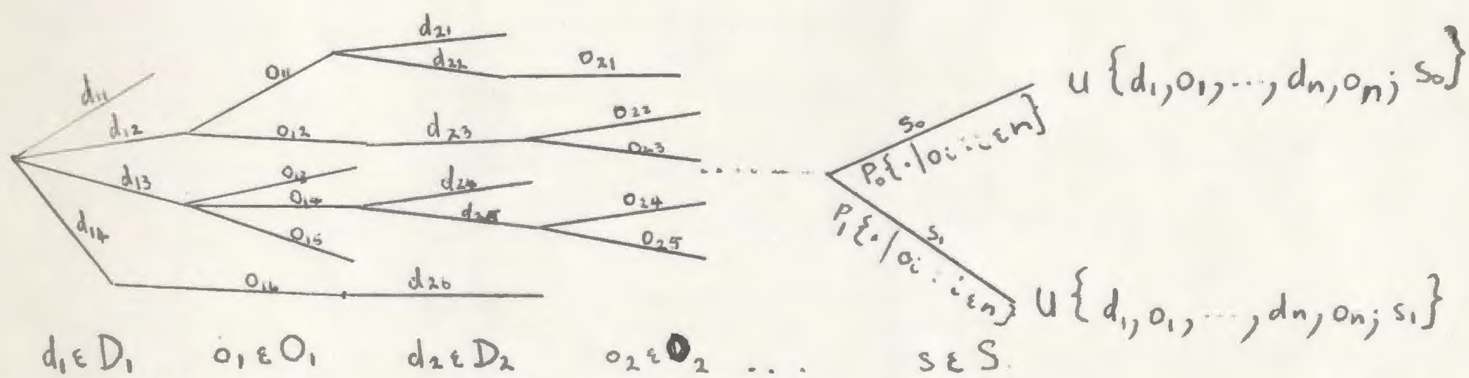
o_2,a_2: Expected profit = (.6)(+310,000) + (.4)(-110,000) = +$142,000.

Maximum Expected Utility =(.5)(-$10,000) + (.3)(+16,000) + (.2)(+$142,000) = +$28,200.

Decisions: Take seismic readings and (i) if no structure is revealed, sell drilling rights, (ii) if open structure or closed structure is revealed, drill, retaining 100% of drilling rights.

We now proceed to the general construction, and we suppose that the petroleum operator has to make a chain of decisions D_i, each having a set of outcomes O_i. Let S be the set of states of the world, and let (d_{ij}) be the set of decisions belonging to each D_i, and (o_{ij}) the set of outcomes belonging to each O_i; then we wish to maximise the expected utility $U(d_1, o_1, d_2, o_2, \ldots, s)$. We achieve this aim by using Bayesian principles, as in examples 1,2, and 3. We then calculate the expected return from each decision-outcome chain and choose the one which maximises the return. A typical generalised decision tree is depicted below. $P(o_i)$ and $P(s_i)$ are the probabilities of the outcomes of geological and geophysical tests, and of the state of the hydrocarbon contents of the reservoir rocks.

Generalised Decision Tree.



The decision tree may be extended as down-hole (stratigraphic) information becomes available. The a priori probabilities differ from province to province and each operator has his own success ratio in a given type of area; these probabilities should be applied in preference to the ones given. A feature which often changes the situation is the liklihood of stratigraphic and hydrodynamic traps in the area; so that even if no structure is revealed the probability of finding oil is relatively higher. Another feature which would favourably change the chance of finding oil would be a seepage in the proximity. One advantage of using a priori probabilities is that personal judgement of these factors enters into the calculation of expected utilities. In actual cases, it is better to tackle the problem piecemeal as in examples 1,2, and 3. We have thus decided the question whether to drill or whether to sell the location.

The next question that arises after drilling the exploratory hole is the better decision take between (i) completing the well, and (ii) selling the drilling rights. We assume that all the log tests are taken in the exploratory hole, so that we make an estimate of the amount of petroleum recoverable.The better decision depends of course on the expected return. Dr. Kaufman gives the following function, which he derived empirically by fitting a logarithmic curve to a set of actual figures supplied by William Beard & Co:

Let v be the value of petroleum found, and $u(v)$ be the utility (profit); then $u(v)$ is defined by $u(v) = -263.31 + 22.093 \log_e(v + 150)$. Thus if Beard does not drill but sells the drilling rights, then $u(0)$ is the profit from finding \$0 of petroleum; and $u(0) = -263.31 + 22.093 \log_e (0 + 150) = -263.31 + 263.31 = 0$; ie his pay-off is 0 utiles..

If he drills the exploratory hole and then abandons the project at a cost of say \$33,750 for 0 barrels of oil, then $u(-33{,}750) = -263.31$

$+22.093 \log_e(-33.750 + 150) = -5.63$ utiles.

If Beard completes a well at a cost of \$100,000 dollars and finds x thousand barrels of oil worth \$2200 profit per thousand barrels then $u(2.2x) = -263.31 + 22.093 \log_e(2.2x + 150 - 100)$.

Thus the break-even point is given when $2.2x + 50 = 150$, ie. when $x = 45.454$. Beard will therefore complete the well if he discovers 45,454 or more barrels of recoverable oil.

We have thus given the optimum decision chain for drilling and completing a well in wildcat territory. The next decision we need to take is where to drill; and this question will be answered in chapters 2 and 3.

CHAPTER 2.

WHERE TO DRILL.

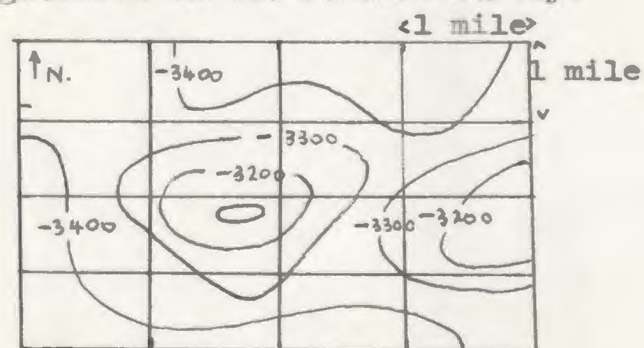
Drilling is the most expensive operation in petroleum exploration, hence the decision regarding the precise location of drill-holes can be vital to a company. This is where all information available must be employed to maximise the expected return.

The sedimentary regions of the earth are places where petroleum deposits have accumulated; these places are depicted in L.G. Weeks' maps in Bulletins of the American Association of Petroleum Geologists,Vol.33 (1949), and vol.49 (1965), the latter being the off-shore sedimentary basins. These off-shore basins are being actively surveyed and explored now; one such region is the Grand Banks off Newfoundland.

The late A.I.Levorsen in his famous work: "Geology of Petroleum" gives the following general information: "Most petroleum has been found in traps that might be classed as either wholly or partly structural. The two most important features of structural traps are the wide variety of structural conditions that may form traps, and the fact that a structural trap may extend vertically through thick sections of potentially productive rocks. Structural mappings have been the most consistently successful method of locating traps. There are several ways of mapping structure - surface, subsurface, core-drill, and geophysical; each of these has as its objective the finding of locally high-structural conditions in underground reservoir rocks that might prove to be traps in which oil or gas or both have accumulated.

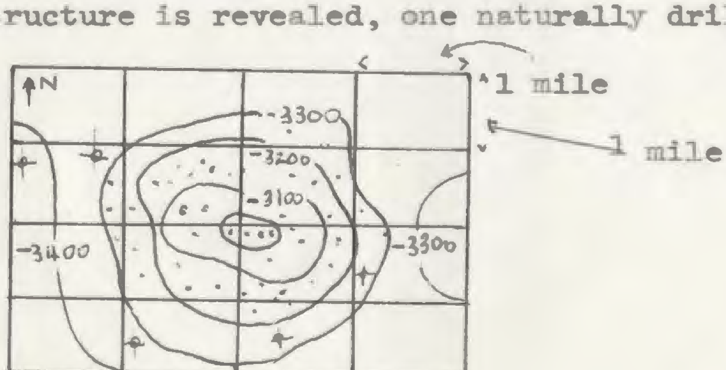
Levorsen further states that: "Where clean, widespread or blanket sands occur, the regional dips are high, and where sloping piezometric surfaces are known, the structural traps generally require a closure to be effective. Where the reservoir rocks are lenticular and variable, minor deformation may be sufficient."

Typical cases are mapped in Chapter VII, Levorsen, and the following example from page 595 gives maps of Paul's Valley field, Garvin County, Oklahoma. Map A_1 is a reconnaissance reflection seismic structural map made in August, 1930, showing the subsurface structure, as construed from the information gathered in the seismic survey.



MAP A_1: Seismic Survey, 1930.

Map B_1 shows the same area after the field was drilled on the discovery of oil in April,1942, the producing sand being Bromide (Ordovician). When a structure is revealed, one naturally drills first at the high points.



Key: . Oil-producer
⊕ Dry hole
MAP B_1: After Drilling, 1942.

Geochemical maps are used for various kinds of chemical analyses of rocks and their fluid contents. Such maps show the surface distribution of hydrocarbons, or waxes, or bacteria which utilize hydrocarbons.Where such a halo is found the inference is that there is a show up and seepage of hydrocarbons from a petroleum reservoir. Soil analysis is used to detect such phenomena. Other geochemical maps are made from data supplied by cores and drill cuttings. Ethane, propane, butane,and pentane, and higher hydrocarbon fractions are measured.

Some oil pools show a significant increase in hydrocarbon content in shales immediately overlaying the reservoir rock, and discoveries of oil-pools have resulted from deeper drilling after encountering shales with a high hydrocarbon content.

The decision on where to drill is made by considering the type of structure revealed in the survey; also by constructing information entropy contours to predict the location with the highest probability of success. In this thesis we shall produce mappings of discriminatory decision functions, with contours which indicate the most likely positions for producing wells, and those which are poorer prospects. Each well will have a discriminant score, which will be the strongest probability index for that control point. This idea will be developed in Chapter 3. In this chapter we shall consider two kinds of information entropies in detail.The first was given by John Dowds in "Computers in Mineral Industries,part 2", a symposium at the School of Earth Sciences, Stanford University,California, 1964.The question, "What is information and how can it be measured?", was the subject of research by Shannon and Wiener, who defined it as follows in "Cybernetics" (Bell). The definition of a measure of information, S, for a continuous function is $S = -\int_{-\infty}^{+\infty} p(x) \log p(x)\, dx$, and for a discrete set is

$S = -kn(\sum_{i=1}^{n} p_i \log p_i + q \log q)$, where the p_i's are the ratios of the producing intervals to the total intervals; q the non-producing

S was called an entropy, since it was thought to be similar to the thermodynamic entropy of statistical mechanics. To simplify the problem Dowds specified that a productive interval is not considered as a function of porosity and saturation $f(\phi, S_w)$; but only that a 10 -ft. interval has a hydrocarbon saturation of 60% or better.

The information we have as we drill an exploratory hole is furnished by drill cuttings and by electric-logs, and radio-active logs passed down-hole; with this information we can work out an entropy value for each hole, and when we have sufficient control points, draw in entropy contour oil-field trends.

In Chapter 13, Levorsen lists the instruments used to measure fluid content, porosity, permeability of rocks and other information vital to the operator.

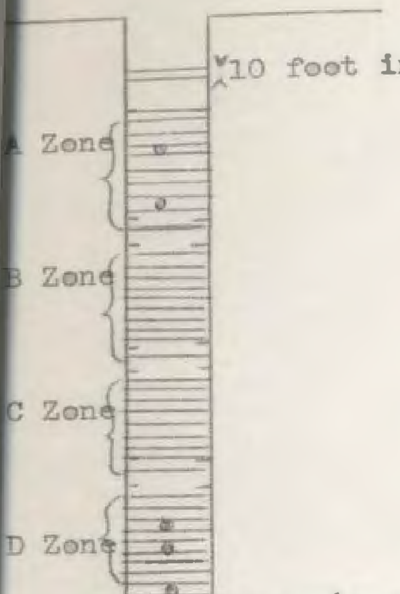
An example will serve to show how an information entropy may be calculated for a given hole (as illustrated). We assume there are 4 strata, A, B, C, and D, which are productive or potentially productive. Each stratum is divided into 9 ten-foot intervals.

EXAMPLE 2.1.

$$\text{Entropy} = kn\left(\sum_{i=1}^{n} p_i \log \frac{1}{p_i} + q \log \frac{1}{q}\right)$$

here n=4, k is arbitrary, say k=90.

so entropy = (90)(4)(2/36 log 18)



$$+ 3/36 \log 12 + 31/36 \log 36/31)$$

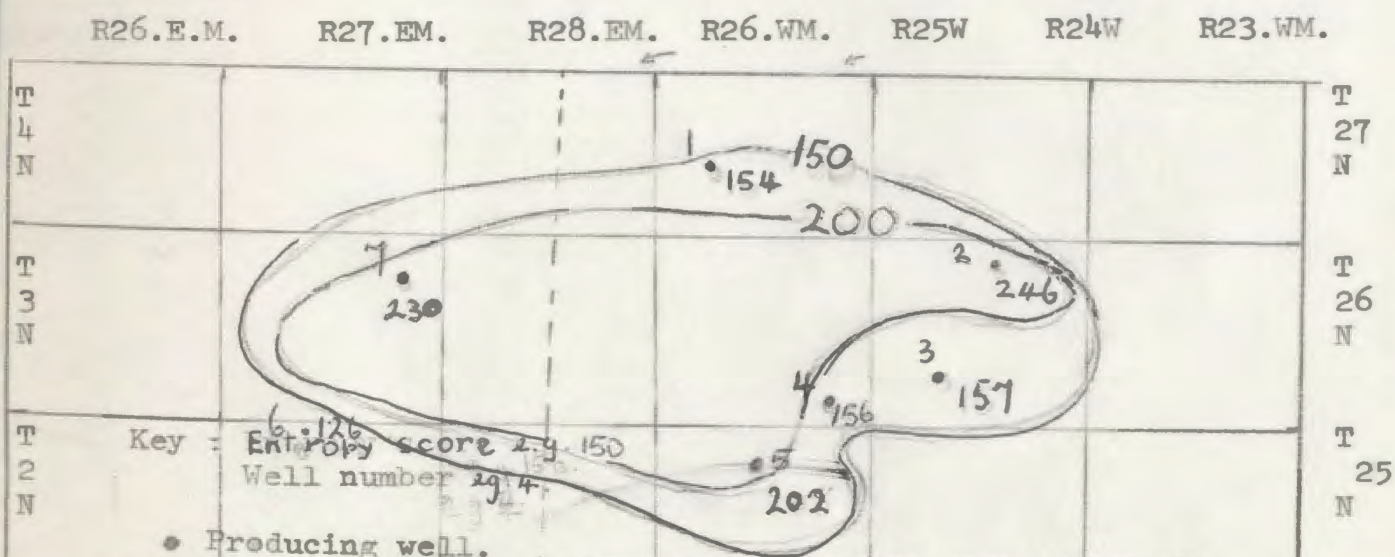
$$= (10)(2.51 + 3.24 + 2.00)(50/36)$$

$$= 107.5$$

$= 108$ to nearest integer.

Hole Data:(2,0,0,3);Entropy 108.

In this way entropy scores may be calculated for each well, and they then serve as probability indices on which we can draw oil-potential regression contours. Dowds gives a case-history of the application of this method to the Mocane-Laverne region of North-West Oklahoma , a 125 square-mile field.The following 7 wells were the only holes in the region in July, 1956 (as illustrated on the map A_2).

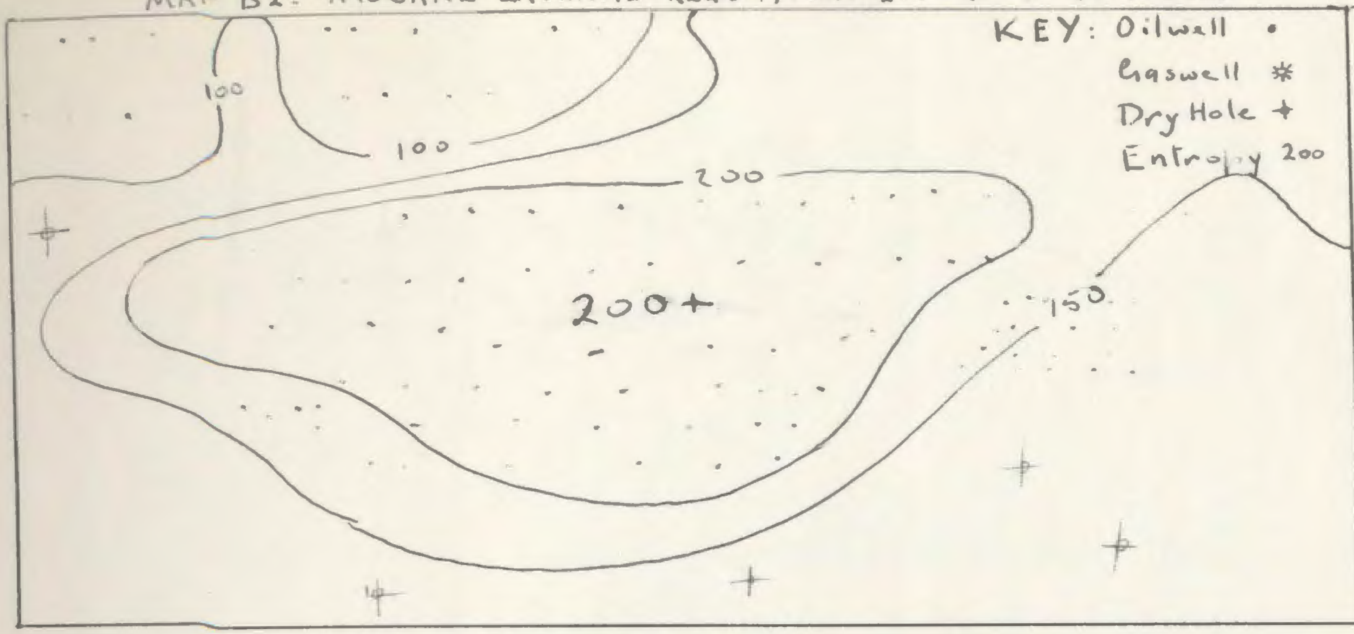


1956: MOCANE LAVERNE REGION, N.W. OKLAHOMA.

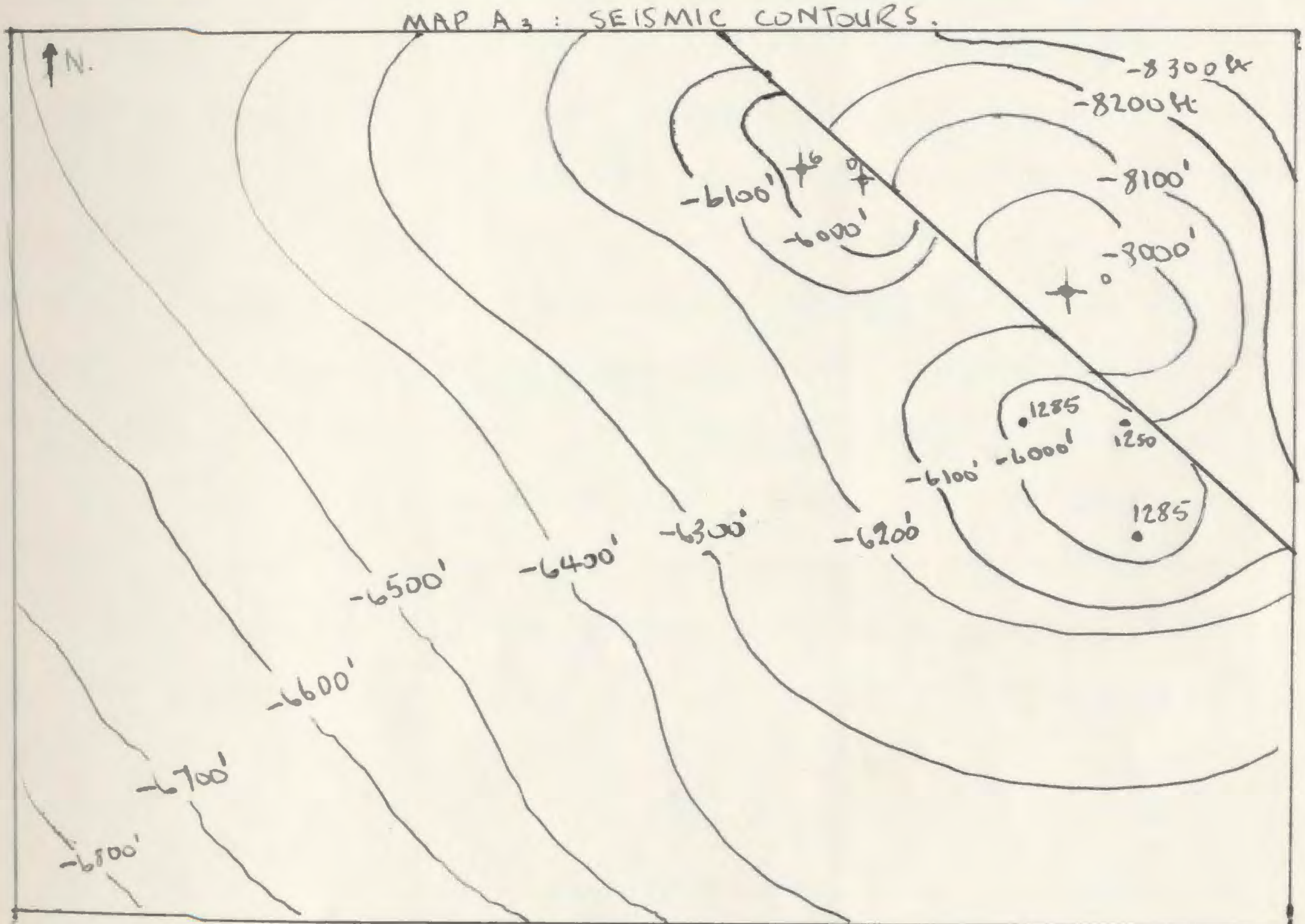
MAP A_2

The data and relevant entropies for the 7 wells are given in the table.

MAP B2: MOCANE-LAVERNE REGION, 1961: ENTROPY CONTOURS.



MAP A3: SEISMIC CONTOURS.



KEY: Oilwell •
Dry Hole +
Seismic Contour ~6000 ft~
Entropy 1285.

Well Number	Data	Entropy
1	(1, 0, 1, 0, 0, 0, 1, 1)	154
2	(1, 1, 1, 0, 0, 1, 1, 1)	246
3	(1, 1, 0, 1, 0, 0, 1, 1)	154
4	(1, 1, 0, 0, 0, 1, 1, 1)	154
5	(1, 1, 0, 1, 0, 1, 1, 1)	208
6	(1, 0, 0, 0, 1, 1, 1, 1)	126
7	(1, 1, 1, 0, 0, 1, 1, 1)	230.

Even with this limited information,(but note 7 producing wells out of 7), tentative oilfield trends could be sketched by using the entropy scores as above. By 1961, it was shown that practically no dry hole could be drilled in the whole 125 square miles. The pay strata which produce oil and gas are from top to bottom: Council Grove of Permian age,Hoover,Toronto,Tonkawa,Lansing-Kansas City, and Morrow of Pennsylvanian age; and finally Chester of Mississippian age.This is an excellent place to study the multiple strata rocks which are productive, and the manner in which hydrocarbons crowded together to form commercial reservoirs.By 1961, the Mocane-Laverne region had over 250 producing well holes drilled, with entropy contours sketched as in figure 2.Calculation of the entropies by thermo-dynamic principles yield similar results, as shown by the following example:

EXAMPLE 2.2. Data:(2, 0, 0, 3); $S = k \log n!/n_1!n_2!..$ by thermo-dynamics

$S = k \log 5!/2!3! = 100 \log 10 = 100$,

compared with 78 by information theory.

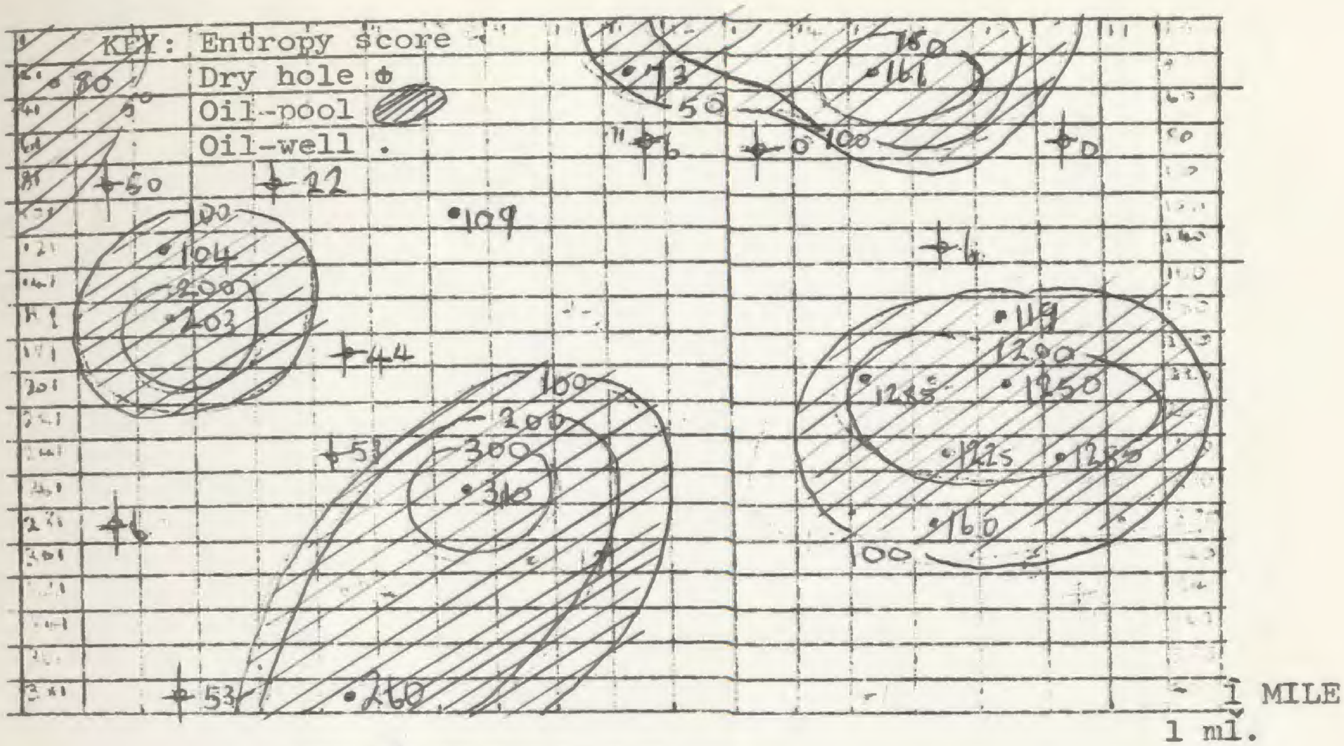
Dowds then gives a hypothetical example in which information entropy is used to discover petroleum. In the map A3), the seismic recordings indicate closed structures and a fault running from N.W. to S.E., with the up side to the West and about a 2000 feet displacement on the down side to the East. Naturally, where we drill first depends upon our knowledge of local conditions; but it will probably be at the higher points in or near to closures. In the model ,hole 71 was drilled first; this was a 'dry' hole with interval values (0,0,1,0), the entropy being S = (100)(1)(1/36 log 36 + 35/36 log 36/35)

= (155.63 + 42.7)/36 = 6 to nearest integer.

The next hole to be drilled was number 73 with data (0,0,0,0) and so the entropy is 0. The next was no.136; data (0,0,1,0), entropy 6. Then number 216; data (0,1,3,6) and thus the entropy =

S =(100)(10)(1/36 log 36) + 3/36 log 12 + 6/36 log 6 + 26/36 log 36/26)

= 1,285. Hole no. 216 is therefore completed as an oil-well; then 218 and 259, all producers. After these locations have been drilled, where do we drill next? Other wildcatters usually move into the area whenever oil is discovered, and drill with or without information; and before long, one has sufficient control points to enable one to sketch in entropy contours for the whole region. In this example, 30 holes were drilled first and the entropy scores are as indicated on map no.2. Entropy contours can now be drawn as illustrated. The oil-pools indicated have been shaded in black. On this basis an additional 52 holes were drilled in this oil-field, at an average cost of $50,000 each, a total of $2,600,000; 70% were successful and reserves worth $13,150,000 were found. Of the original 30, 15 were successful and for an



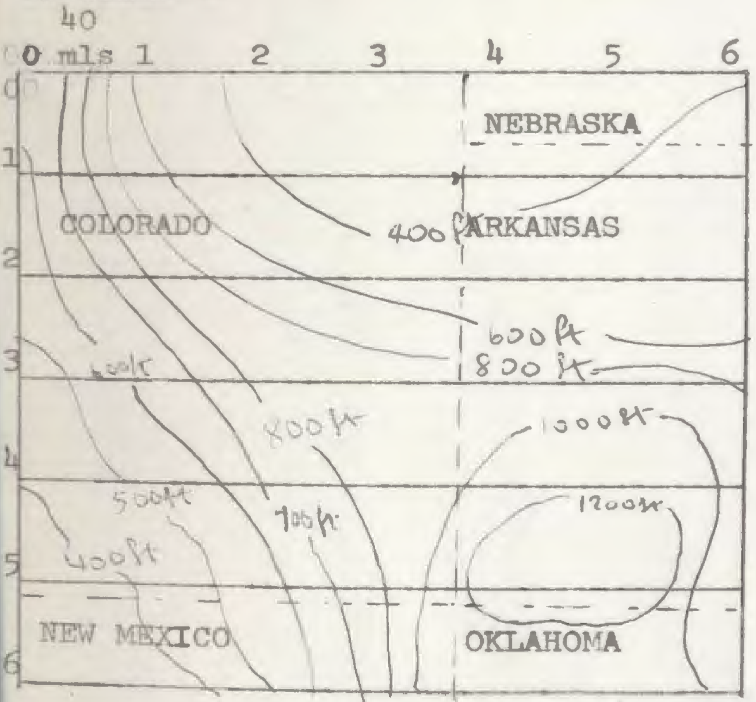
MAP [illegible], ENTROPY CONTOURS.

outlay of \$1,500,000, reserves worth \$5,100,000 were discovered.

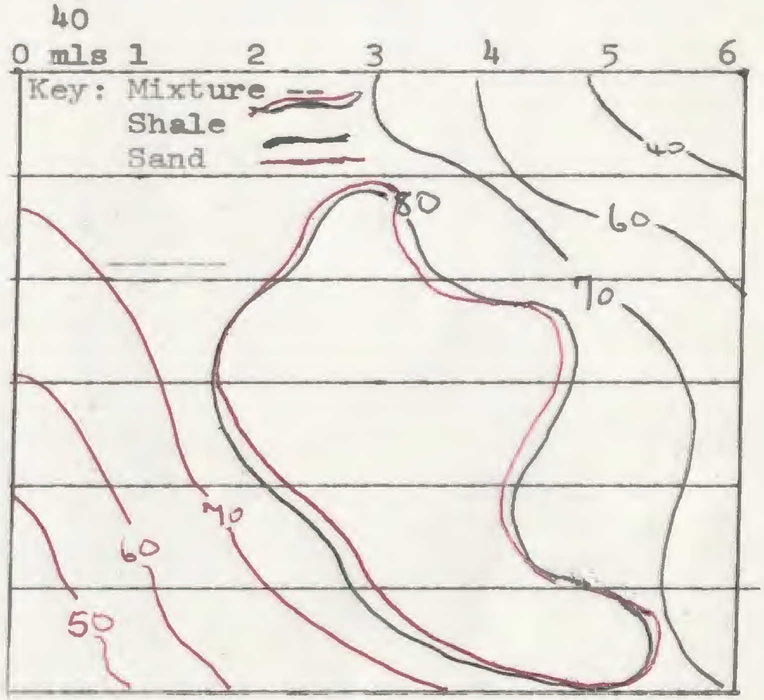
A second set of entropy contours we can employ is directly connected with the way oil deposits have accumulated. We will give an example of its utility.Let p_1 = % of shale, p_2= % of sand-stone, p_3 = % of carbonate and p_4 = % of evaporite; these statistics having been secured from logs and drill cuttings. We define S, the facies entropy, as $S = -1/\log 4 \sum_{i=1}^{4} p_i \log p_i$. We then plot the entropy values on the map and fit the best entropy contours. A facies entropy map is useful basically because it relates to the theory of how hydrocarbons are accumulated. The position of ancient shorelines is important in the generation of hydrocarbon deposits, and in the collection and deposition of porous rocks. All of them are related to energy, and the point of maximum energy is where air and water met: on the shoreline.

Here, there is a maximum mixing of the four elements. Sometimes this mixing forms a halo effect around the areas most favourable for hydrocarbon deposits, oil deposits usually occurring in areas of low entropy mixing, where the mixture clears up to sand in a fairly small areal unit such as a bed or horizon in a producing region. Thus the areas with the lowest facies entropy associated with thick strata, and high producing-potential entropy are the places to drill first.

The isopach map A_4 and entropy facies map B_4 were prepared from data given in "Stratigraphic Mapping" by Dr. Krumbein, in the journal of the American Association of Petroleum Geologists, 1962; reproduced in table C.



MAP A_4 ISOPACH.

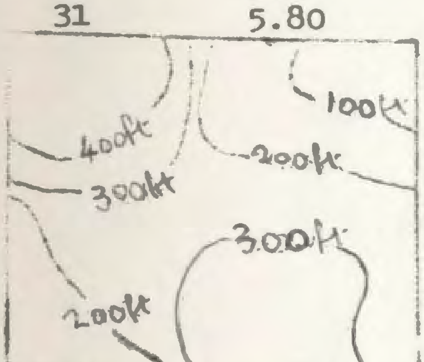


MAP B_4 FACIES ENTROPIES.

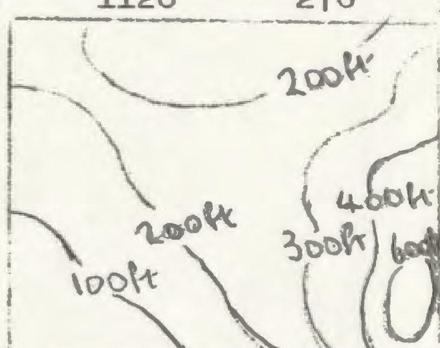
Maps C, D and E depict a sand isolith, shale isolith, and sand-shale ratios for the same region.

TABLE C: COORDINATES AND THICKNESSES IN FEET OF THE FOUR COMPONENTS IN THE PERMIAN BASIN,DENVER,COLORADO.

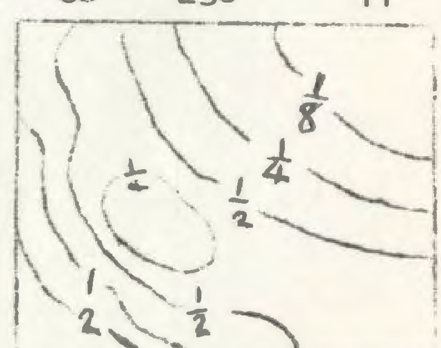
Control Point.	U Coord.	V Coord.	Total Thickness	Sand	Shale	Carbon -ate	Evapor -ite	Facies Entropy
1	2.60	1.85	608	365	148	20	75	74
2	2.85	2.35	640	224	304	14	98	79
3	2.30	2.60	464	104	242	18	100	82
4	2.20	4.50	532	157	238	0	137	77
5	2.30	5.50	562	120	316	0	126	71
6	1.40	5.55	530	30	461	0	39	34
7	2.95	0.20	447	293	116	12	26	64
8	3.30	1.15	844	451	311	42	40	72
9	3.40	2.30	906	337	432	60	77	80
10	3.55	3.10	845	266	350	24	205	88
11	3.80	2.90	915	295	355	43	222	88
12	4.00	3.60	1139	179	643	20	297	75
13	3.65	3.70	1118	180	568	0	370	72
14	4.20	3.85	1224	207	758	11	248	76
15	3.45	4.80	1162	130	659	13	360	71
16	3.30	5.10	1003	224	542	21	216	78
17	3.10	5.55	721	229	400	12	80	73
18	3.00	6.20	775	223	477	28	47	68
19	4.35	0.60	374	240	110	24	0	59
20	4.30	1.15	614	255	272	28	59	79
21	4.95	2.25	702	237	341	39	85	82
22	5.00	2.60	933	275	435	41	182	85
23	4.85	3.10	1001	348	450	17	186	80
24	4.40	4.26	1204	277	610	10	301	77
25	5.10	4.10	1144	310	520	12	302	80
26	5.50	3.80	1048	362	510	12	164	76
27	5.30	4.30	1114	246	528	32	308	83
28	5.50	4.20	1023	295	501	18	209	80
29	4.60	5.70	955	267	502	24	162	78
30	5.10	5.75	1005	271	637	8	89	76
31	5.80	3.40	1126	270	558	68	230	77



MAP C4 SAND ISOLITH



MAP D4 SHALE ISOLITH



MAP E4 SAND-SHALE RATIO

The next important mapping we consider is the mapping of principal component scores, which will provide a powerful tool in petroleum exploration. Firstly, however, we will consider principal component analysis.

Principal components are linear combinations of statistical variables, which have special properties in terms of the variances. The first principal component, with which we shall be concerned in this thesis, is defined as the normalised linear combination of the variables such that it has the maximum variance. Thus if $y=\sum_{i=1}^{n} a_i x_i$ is the principal component, where the x_i are the data measured from their respective means, then the a_i are calculated to maximise the variance of y. The principal components turn out to be the characteristic vectors of the covariance matrix. Thus the study of principal components can be considered as the statistical development of characteristic roots. In effect, transforming the original vector variable to the vector of principal components amounts to a rotation of coordinate axes to a new coordinate system. Andersen gives the following definition:

Definition of Principal Components: Let S be the covariance matrix of the vector $X= (x_1, x_2, \ldots, x_n)$ and let the mean vector = 0. Let B be an n-component column vector such that $B'B = 1$. The variance of $B'X$ is $E(B'X)^2 = E(B'XX'B)= B'SB$....(1) where E is the expected value operator. To determine the normalised linear combination $B'X$ with maximum variance, we must find a vector B satisfying $B'B = 1$ which maximises (1). Let $\phi = B'SB - L(B'B - 1) = \sum_{ij} B_i s_{ij} B_j - L(\sum_i B_i^2 - 1)$, where L is the Lagrange multiplier.

The vector of partial derivatives $d\emptyset/dB_i$ is $d\emptyset/dB = 2SB - 2LB$...(2), hence a vector B maximising BSB' must satisfy expression (2) equated to 0, so that:

$(S - LI)B = 0$.....(3).

In order to get a solution of (3) with $B'B = 1$ we must have $S - LI$ singular, so that L must satisfy $|S - LI| = 0$...(4); the left side of (4) is a polynomial of degree n in L; so that (4) has n roots; let these be $L_1 = L_2 = L_3 = \ldots = L_n$. If we multiply (3) on the left by B' we obtain

$B'SB = LB'B = L$, showing that B satisfies (3)

and $B'B=1$, thus the variance of B'X is L; and so for a maximum variance we choose L_1, the largest L. Let A be the normalised solution of $(S - L_1 I)B = 0$, then $Y = A'X$ is the normalised linear combination with maximum variance. This is the first principal component. We shall employ an approximate method for finding L_1 the largest characteristic root using Kendall's method from "A Course in Multivariate Analysis" (Harper, N.Y.). We now apply this analysis to find the principal components, y_i, accounting for the accumulation of petroleum. We calculate table D from table C to give the following correlation matrix.

TABLE D: CORRELATION MATRIX from data of TABLE C.

	X_1 Total Thickness	X_2 Sand %	X_3 Shale %	X_4 Carbonate %	X_5 Evaporite %
X_1 Total Thickness	1	-.73	+.56	-.48	+.67
X_2 Sand %	-.73	1	-.85	+.44	-.76
X_3 Shale %	+.56	-.85	1	-.42	+.33
X_4 Carbonate %	-.48	+.44	-.42	1	-.46
X_5 Evaporite %	+.67	-.76	+.33	-.46	1

We now give Kendall's method for finding $y_1 = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5$, the first principal component.

Adding the columns of the correlation matrix, we have totals = (1.98, -.80, +.62, +.08, +.58), and dividing by the supremum gives: (1, -.40, +.31, +.04, +.29). Write this row as column A below and multiply the rows of the original matrix by the corresponding row of A,

A	The new matrix is:					B
1	1	-.73	+.56	-.48	+.67	1
-.4	+.29	-.40	+.34	-.18	+.30	-.98
+.31	+.17	-.26	+.31	-.13	+.10	+.79
+.04	-.02	+.02	-.02	+.04	-.02	-.54
+.29	+.20	-.22	+.10	-.14	+.29	+.82
Totals =	1.64	-1.59	1.29	-.89	1.34	

Repeating the process gives column B, and iterating we have:

					C
1	-.73	+.56	-.48	+.67	1
-.29	+.40	-.34	+.18	-.30	-.71
+.13	-.20	+.24	-.10	+.08	+.54
+.01	-.01	+.01	-.02	+.01	-.52
+.16	-.18	+.08	-.11	+.24	+.73
1.01	-.72	+.55	-.53	+.73	

5 more iterations gives $(a_1, a_2, a_3, a_4, a_5)$ proportional to (1, -.7, .5, -.5, .7) so that $L_1 = (1)(1) + (-.7)(-.73) + (.5)(.56) + (-.5)(-.48) + (.7)(.67)$ = approxim 2.57, thus the first principal component accounts for 2.57/5 x 100% = 51.4% of the variation in the accumulation of petroleum.; and $y_1 = \dfrac{x_1 - .7x_2 + .5x_3 - .5x_4 + .7x_5}{\sqrt{(.7^2+1^2+.5^2+.5^2+.7^2)}}$

Therefore $y_1 = .63\,x_1 - .44\,x_2 + .32\,x_3 - .32\,x_4 + .44\,x_5$.

To find the 2nd principal component we form the matrix $L_1 a_i a_j$ =

1	-.7	+.5	-.5	+.7
-.7	+.49	-.35	+.35	-.49
+.5	-.35	+.25	-.25	+.35
-.5	+.35	-.25	+.25	-.35
+.7	-.49	+.35	-.35	+.49

Subtracting this from the original matrix, we get the residual matrix:

					A
0	-.03	+.06	+.02	-.03	+04
-.03	+.51	-.50	+.09	-.27	-.34
+.06	-.50	-.75	-.17	-.02	+.19
+.02	+.09	-.17	+.75	-.11	1.00
-.03	-.27	-.02	-.11	+.51	+.14
+.02	-.20	+.11	+.58	+.08	

Iterating, we have:

0	0	0	0	0
.01	-.17	+.17	-.03	+.09
.01	-.10	+.15	-.03	0
.02	+.09	-.17	+.75	-.11
0	-.04	0	-.07	+.o7
04	- 22	15	.67	+.05

B	C	D
.06	.03	.03
-.33	.17	.12
+.22	-.26	-.25
1.00	1.00	1.00
.07	-.18	-.16

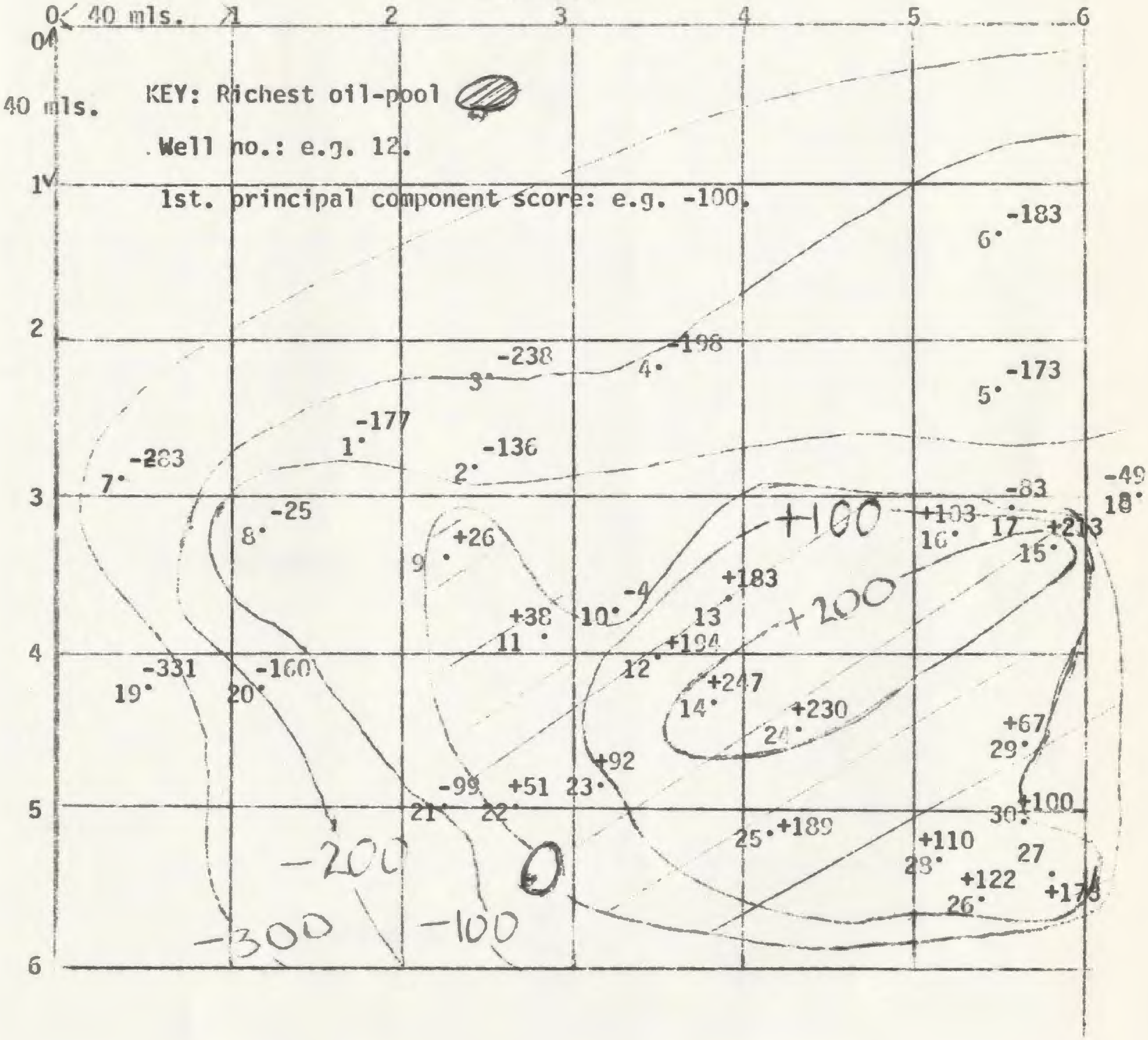
After two more iterations, the columns become identical (to within two d.p.) Thus $(a_1,a_2,a_3,a_4,a_5) = (.02,.09,-.17,.75,-.11)/\sqrt{(.612)}$; so that y_2, the second principal component is given by :

$y_2 = .03x_1 + .11x_2 - .21x_3 + .94x_4 - .14x_5$, and $L_2 = .612$; thus the second principal component accounts for .612/5 = 12.2% of the variation in oil deposits. Since the L_i's are in descending order and the other three components account for 36.4% of the variation, each must = 12% approximately. Therefore the first principal component is the only outstanding one, and it must therefore be a very powerful exploration tool. Table E below gives the first principal component scores for each well; and map F overleaf gives the mapping indicating that the richer petroleum deposits lie in the South-East, the deposits becoming poorer to the North and West. This prediction is most important when one considers each square is 40 mls X 40 mls.

TABLE E: FIRST PRINCIPAL COMPONENT SCORES IN THE PERMIAN BASIN, DENVER, COLORADO.

WELL NO:	1	2	3	4	5	6	7	8	9	10
PRINCIPAL COMPONENT:	-177,	-136,	-239,	-197,	-172,	-183,	-283,	-25,	26,	-4
WELL NO:	11	12	13	14	15	16	17	18	19	20
PRINCIPAL COMPONENT:	38,	194,	183,	247,	213,	103,	-83,	-49,	-331,	-160,
WELL NO:	21	22	23	24	25	26	27	28	29	30
PRINCIPAL COMPONENT:	-99,	51,	92,	230,	189,	122,	173,	119,	67,	100.

MAP F4: 1ST. PRINCIPAL COMPONENT IN THE PERMIAN BASIN, DENVER, COLORADO.



We will defer discussion of the most useful mapping, discriminant scores,until the last chapter, and we return to the question of the best regions of exploration.In his dissertation, "Analysis of Petroleum Potential through regional geological Synthesis", in the Bulletin of the American Association of Petroleum Geologists,1963, W.W. Mallory summatizes the barren and producing sedimentary rocks of North America. He compares the Leduc oil-pool of Southern Alberta with the Norman Wells pool in the District of MacKenzie over 1,000 miles to the North, by the fact that each is producing from the same cratonic sedimentary rocks, comprising a sequence of Devonian Reefs. Levorsen stated that the Devonian reefs of Western Canada may constitute a reserve far greater than at present suspected. In a similar way, Mallory compares the Western coast of Newfoundland and the off-shore region with Stony Creek oil and gas pool in the Moncton basin, which is producing at between 5-10,000 feet; and also with the Gaspé peninsula with its many oil seepages, based on the continuity of the Mississippian and Pennsylvanian rocks which underlay the whole region. Regional geological synthesis is therefore a most useful tool in the search for petroleum deposits in unexplored territories. As drilling progresses, and subsurface data becomes available,regional synthesis and discriminatory analysis provide the most reliable contours for extrapolating from developed petroleum reserves to the ultimate potential of a region. Specifically we often infer that the thinning out of sandstone to an impermeable rock is a stratigraphic trap. Mallory then maps and classifies the whole North American Continent according to its geology related in the following way to its petroleum potential:

NONE: 1. Precambrian rocks of the Canadian Shield.

2. Precambrian basement rocks beneath cratonic sedimentary strata.

3. Tertiary Volcanic rocks.

POOR: 1. Eugeosynclines, with the possible exception of the Gaspé Peninsula.

2. Miogeosynclines, except for the Craton-marginal belts of Alberta,District of MacKenzie, Oklahoma, and W. Virginia.

GOOD TO EXCELLENT:

1. Gulf-Atlantic coastal plain:Tertiary, Cretaceous, and Triassic systems.

2. Arctic coastal plain.

3. Basins developed on old eugeosynclines,especially Tertiary rocks of S.California; possibly E. Canada coastal region and Western Newfoundland.

4. Craton, and all Cambrian and younger strata..

Mallory's map then indicates the classification into the above categories of the whole North American continent. This could give a more accurate estimate for the a priori probabilities in a given location, and could be applied in our discriminatory decision function defined in Chapter 3.

CHAPTER 3.

DRILLING DECISION DISCRIMINATORY FUNCTION

In any particular location there are two states of the world: (subsurface hydrocarbon content): (1) the well if drilled does not contain oil in commercial quantities. We shall say that wells in this category belong to population (1), the population of non-producers. (2) The well does contain oil in commercial quantities. Wells in this category we shall say belong to population (2), the population of producers.

We should like to formulate discriminatory decision functions, which will assign each location to its correct population (1) or (2). We shall do so using Bayesian a priori probabilities, and two sets of data: one from a set of wells belonging to population (1); and the other from a set belonging to population (2). We will now consider our discriminatory analysis to formulate suitable functions.

To discriminate at all, we must have data on a set of wells from each population; with these data, we can then define a function which will be the most powerful discriminator, and furthermore should be the best function for determining the measures which lead to producers and those which lead to non-producers.We shall consider three discriminant functions; the first two by Fisher and Anderson are equivalent when classification into two populations is required.

R.A. Fisher defined the first linear discriminating function, the L.D.F. He gave the answer to the question: what linear function of the well data $x= (x_1, x_2, \ldots, x_p)$, $X = b_1x_1 + b_2x_2 + \ldots + b_px_p$, will maximize the ratio of the difference between the specific means of the two populations to the standard deviation within the species (the whole set of wells taken together). Let the differences between the means be $(d_1, d_2, \ldots, d_p)$, then for the linear function defined, the difference between the means of X for the two populations is:

$$D = b_1 d_1 + b_2 d_2 + \ldots + b_p d_p,$$

whilst the variance within the species is proportional to:

$$S = \sum_{i=1}^{p} \sum_{j=1}^{p} b_i b_j S_{ij} \text{ where } S_{ij} \text{ is the variance or covariance.}$$

The linear function which best discriminates between the two populations is the one for which the ratio D^2/S is a maximum, by independent variation of the coefficients b_i, $i = 1,2,\ldots,p$. This gives for each

$$b_i \quad D/S^2(\ 2S\ dD/db_i - D\ dS/db_i) = 0$$

so that $\frac{1}{2}\ dS/db_i = (S/D) dD/db_i$.

Now S/D is a factor constant for all the p unknown coefficients, so that the coefficients required are proportional to the solution of the equations

$$\begin{aligned} S_{11}b_1 + S_{12}b_2 + \ldots + S_{1p}b_p &= d_1 \\ S_{12}b_1 + \ldots\ldots\ldots\ldots + S_{2p}b_p &= d_2 \quad \text{since } S_{12} = S_{21}, \text{ etc.} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ S_{1p}b_1 + \ldots\ldots\ldots\ldots + S_{pp}b_p &= d_p\ . \end{aligned}$$

which may be written more neatly in matrix notation as SB=D.

Thus solving by inverting the matrix S, we have $b_i = \sum_{j=1}^{p} S^{ij} d_j$ where S^{ij} are the terms of the inverse of S, and $i = 1,2,\ldots,p$.

The matrix S may be inverted provided it is non-singular. If it is singular, then the variables x_i are not independent and one or more is a linear function of some of the others. Since these variables add no new information, they can be eliminated, and the discriminatory analysis can be carried out on the remaining variables. We therefore need to discover which of original p variables may be expressed as a linear combination of others. This is determined by the multiple correlation coefficient. If for example x_1 can be expressed as a linear function of the other p-1 variables, then the multiple correlation coefficient, $r_i = 1$. The closer to 1, the better the fit of the regression plane of x_1 on the other variables. Now the multiple correlation coefficient, r_i, of x_i on the other variables is given by: $r_i^2 = 1 - |R / R_i|$ where R is the covariance or correlation matrix, and Morrison in "Multivariate Statistical Methods", page 93, shows that this is equivalent to : $r_i^2 = 1 - 1/ s_i^2 S^{ii}$ where S^{ii} is the diagonal element of the inverse of the covariance matrix.

Hence variables making little or no contribution to the discrimination may be eliminated. Thus if there are q remaining variables, which are linearly independant, the discriminant function may be defined on them as follows: $X = \sum_{i=1}^{q} b_i x_i$, where $b_i = \sum_{j=1}^{q} S^{ij} d_j$; S^{ij} are the terms in the inverse of matrix S, and d_j are the differences between the means of the populations, as before.

The ratio of half the difference between the means to its standard error is of interest in relation to the probability of misclassification of a new well; and also to the fiducial limits which obtain when we assign a new well using the discriminant function. The standard error will also be of importance in determining the minimum number of wells needed before a region is under control;that is before we can set up a discriminant function which is sufficiently reliable.

We shall first require an estimate for the variance of the discriminant function based upon the data of the control points (wells); and we shall call the discriminant Y in this section and S_y the estimated variance of Y. Now $Y = \sum_{i=1}^{p} b_i x_i$, or in matrix terms $Y = BX'$ where B is the row matrix $(b_1,\ldots,b_p)$ and X' is the column matrix $(x_i : i = 1,\ldots,p)$.

Then S_y may be estimated as follows: $S_y = BSB' / (n_1 + n_2 - 2 - p + 1)$, where S is the variance-covariance matrix, and B' is the transpose of B, and where the denominator is the number of degrees of freedom, made up as follows: n_1, n_2 are the number of wells in populations (1), (2) in the sample, less 2 since we used the two means in calculating S, and less p-1 since there are p-1 adjusted ratios in the p coefficients contained in B. The square root of S_y will thus be the standard error of the discriminant score of individual wells.The following theorem yields a simple method of calculating S_y, which we shall use in our field studies:

THEOREM 3.1: $S_y = \{\overline{Y}_1 - \overline{Y}_2\} / (n_1 + n_2 - 2 - p + 1)$ where S_y is the estimated variance of the discriminant scores, and $\overline{Y}_1$, $\overline{Y}_2$ are the sample means of populations (1),(2).

PROOF: $Y = BX'$ where B is the row matrix $(b_1,\ldots,b_p)$ and X' is the column matrix $(x_i : i = 1,\ldots,p)$.

Now $b_i = \sum_{j=1}^{p} S^{ij} d_j$; and hence $B' = S^{-1}D'$, so that its transpose B must $= D(S^{-1})' = DS^{-1}$ since S^{-1} is a symmetrical matrix.

Now, $S_y = BSB' / (n_1 + n_2 - 2 - p + 1)$, and $BSB' = DS^{-1}S\,S^{-1}D' =$
$DSD' = BD' = \overline{Y}_1 - \overline{Y}_2$, so that $S_y = (\overline{Y}_1 - \overline{Y}_2)/(n_1+n_2-2-p+1)$. QED.

The ratio of the difference between the means to its standard error in individual wells is of interest in that it will help us in deciding how many control points (wells) we require to set up a sufficiently strong discriminant function, that is before the region is under control. Supposing a well is misclassified if its deviation from the population mean exceeds half the difference between the means of the two populations, then again, assuming normal distribution for X, we can be 95% confident this will not happen if $1.96\sqrt{S_x}$ does not exceed half the difference between the means. Thus the number of control points required for sufficiently accurate discrimination may be determined.

Anderson used a different approach to the classification problem in his work: "An Introduction to Multivariate Statistical Analysis", and his discriminant is sharper in that it uses the Bayes procedure which takes into account a priori probabilities.He defined the best classification procedure as the one which minimized the average cost of misclassifications; and thus maximised the expected utility of classification of wells. Anderson's theorem states:

THEOREM 3.2: Assign a well to population (1) if $c_1q_1f_1(x) > c_2q_2f_2(x)$ and to population (2) if $c_2q_2f_2(x) \geq c_1q_1f_1(x)$, where $x = (x_1,x_2,...,x_p)$ is the vector data of the well; f_1,f_2 are the respective frequency distributions of populations (1),(2); c_1 is the cost of classifying a well actually a member of population (1) as coming from population (2), and c_2 is similarly defined; and q_1 is the relative frequency of wells belonging to population (1),that is relative to those belonging to population (2), and q_2 is the relative frequency with which wells belonging to population (2) occur. This classification procedure maximises the expected utility, by minimising the expected loss.

To prove this theorem, we shall require the following lemma:

LEMMA.3.1: The average loss from costs of misclassification is:

$L = c_1 q_1 \int_{R_2} f_1(x) + c_2 q_2 \int_{R_1} f_2(x)dx$, where R_1 is the region of classification as from population (1), and R_2 as from population (2).

Proof: Since the probability of drilling a well from population (1) is q_1; and the probability of drilling a well from population (1) and correctly classifying it is $q_1 \int_{R_1} f_1(x)\, dx$, therefore the probability of drilling a well from population (1) and misclassifying it is $q_1 \int_{R_2} f_1(x)dx$. Similarly the probability of drilling a well from population (2) and misclassifying it is $q_2 \int_{R_1} f_2(x)dx$. Therefore the expected or average loss from the costs of misclassification is the sum of the products of the costs of each misclassification times the probability of its occurrence, it is $L = c_1 q_1 \int_{R_2} f_1(x)dx + c_2 q_2 \int_{R_1} f_2(x)dx$. Q.E.D.

It is this average loss we wish to minimise. That is we wish to divide the space into regions R_1, R_2 such that the expected loss is minimised. A procedure which minimises L is called a Bayes procedure. We can now prove theorem 3.2.

THEOREM 3.2. Proof: The probability that a well belongs to population (1), and that each variate is less than the corresponding component in X is $\int_{-\infty}^{x_p} \int_{-\infty}^{x_{p-1}} \dots \int_{-\infty}^{x_1} q_1 f_1(x)dx_1 dx_2 \dots dx_p$, and the conditional probability of a well belonging to population (1) is $q_1 f_1(x) / (q_1 f_1(x) + q_2 f_2(x))$, and to population (2) is $q_2f_2(x) / (q_1f_1(x) + q_2f_2(x))$. Now from lemma 3.1 the expected loss is L, and for a given well, x, we minimise the probability of a misclassification by assigning it to the population with the higher conditional probability, hence we assign it to population (1) if $q_1f_1(x) > q_2f_2(x)$ and to population (2) if $q_2f_2(x) \geq q_1f_1(x)$, the well being assigned arbitrarily to population (2) if the two are equal from wider economic considerations. Similarly to minimise the cost of misclassification we assign a well to population (1) if $c_1q_1f_1(x) > c_2q_2f_2(x)$ and to population (2) if $c_2q_2f_2(x) \geq c_1q_1f_1(x)$.

Q.E.D.

THEOREM 3.3: The Bayes procedure is the optimum procedure.

Proof: For any procedure $R^* = (R_1^*, R_2^*)$, the probability, y, of misclassification is, by an argument similar to the one given in the last proof;

$$y = q_1\int_{R_2^*} f_1(x)dx + q_2\int_{R_1^*} f_2(x)dx$$
$$= \int_{R_2^*}(q_1f_1(x) - q_2f_2(x))dx + q_2\int_{R^*} f_2(x)dx.$$

Now the q_i and f_i are non-negative, thus R_2^* includes the points, x, such that $q_2f_2(x) \geq q_1f_1(x)$ and excludes the points such that $q_1f_1(x) > q_2f_2(x)$, thus the Bayes procedure is unique.

We note that, mathematically, the problem was: given non-negative constants q_1, q_2 and non-negative functions $f_1(x)$, $f_2(x)$, choose regions R_1, R_2 to minimise $c_1q_1\int_{R_2} f_1(x)dx + c_2q_2\int_{R_1} f_2(x)dx$, thus we choose R_1, R_2 such that R_1: $c_1q_1f_1(x) > c_2q_2f_2(x)$ since c_1, c_2 are non-negative constants, and R_2: $c_2q_2f_2(x) \geq c_1q_1f_1(x)$.

The Bayes procedure is thus the optimum procedure.

Anderson's next theorem deals with the best classification procedure, when the two populations are assumed to be normally distributed; we shall see that apart from the Bayesian a priori probabilities, the Anderson discriminant function turns out to be identical to Fisher's L.D.F.

THEOREM 3.4: Let (m_1,S), (m_2,S) be the parameters of Populations (1),(2) respectively, and let c_1,c_2,q_1,q_2 be the probabilities and costs previously defined, then assuming the populations to be normally distributed with a common covariance matrix S, the following classification procedure will maximise the expected utility:

Assign a well ,x, to population (1) if $x'S^{-1}(m_1-m_2) - \frac{1}{2}(m_1+m_2)'S^{-1}(m_1-m_2) > \log_e\left(\frac{c_2q_2}{c_1q_1}\right)$, and to population (2) if it is less than or equal to $\log_e\left(\frac{c_2q_2}{c_1q_1}\right)$, where m_1, m_2 are the mean vectors of the two populations.

Proof: From theorem 3.2, we classify a well,x, as belonging to population(1), if $c_1q_1f_1(x) > c_2q_2f_2(x)$; otherwise we classify it as population (2).

Now since $f_1(x), f_2(x)$ are assumed to be multivariate normally distributed we have $f_i(x) = (1/\left[(2\pi)^{\frac{1}{2}p}|S|^{\frac{1}{2}}\right]\left\{\exp\left[-\frac{1}{2}(x-m_i)'S^{-1}(x-m_i)\right]\right\}$, i=1,2.

Therefore, $f_1(x)/f_2(x) = \exp(-\frac{1}{2}(x-m_1)'S^{-1}(x-m_1)) - -\frac{1}{2}(x-m_2)'S^{-1}(x-m_2))$

$=\exp(\frac{1}{2}(x)'\ S^{-1}(m_1-m_2) - \frac{1}{2}(m_1+m_2)'\ S^{-1}(m_1-m_2))$

Taking logarithms to base e in the classification inequality from theorem 3.2, we assign a well,x, to population (1) if $\log_e f_1(x)/f_2(x) > \log_e c_2q_2/c_1q_1$; that is if: $x'\ S^{-1}(m_1-m_2) - \frac{1}{2}(m_1+m_2)'S^{-1}(m_1-m_2) > \log_e c_2q_2/c_1q_1$, otherwise we assign the well to population (2).

Q.E.D.

The first term on the left side of the inequality is Fisher's L.D.F, and the second term is the discriminant score of the well mid-way between the means of the two populations. In actual examples, we shall therefore use

Fisher's Linear discriminant function together with the Bayesian a priori probabilities applying in the oilfield where the wells are being drilled, to assign a new well to its correct population. In general we shall take c_1 to be \$100,000, c_2= \$300,000 ; q_1= .9 and q_2= .1; so that we shall assign a location to population (1) if $X - \overline{X}$ exceeds $\log_e 1/3 = -1.1$, and to population (2) if $X - \overline{X} = - 1.1$, where X is the discriminant score of the well, $\overline{X}$ is the discriminant score of the well mid-way between the means of the two populations. The discriminant scores will also be extremely useful for mappings for predicting the locations of new oil wells. We now give an actual field example to demonstrate the power of the discriminant function, both in discriminating between non-producers and producing wells; and in predicting the best locations to drill next.

I have called the oil-field the Sproule Oil-field as Messrs Sproule & Associates supplied the data of the 22 wells at present drilled. Information as to its actual location can be supplied by Sproule of Calgary,Alberta.

FIELD STUDY 3.1: Sproule Field, Alberta; Twp...,Rge 7, W4M.

TABLE 3.1: WELL DATA

a) PRODUCING WELLS:

Well Number LSD.	Section	x_1:feet sub-sea:VIKING.	X_2:feet sub-sea:BL.	x_3:feet sub-sea:BSL QTZ.
2	21	67	-114	-437
6	21	65	-105	-435
7	21	58	-109	-431
8	21	59	-105	-440
11	21	52	-112	-443
6	27	63	-109	-439

WELL NUMBER.		x_1	x_2	x_3
11	27	49	-121	-455
12	27	24	-113	-448
15	28	46	-121	-442
4	29	46	-114	-443
2	33	39	-125	-459
4	33	41	-123	-449
NON-PRODUCERS				
5	1	37	-129	-533
5	4	31	-146	-525
1	22	60	-113	-433
4	22	47	-124	-490
13	24	44	-136	-619
4	27	46	-116	-483
9	33	31	-130	-455
10	33	-	-129	-478
16	33	42	-128	-467
4	35	48	-126	-520

To find the linear discriminating function we first need to find S, the variance-covariance matrix. Now the mean vector of the 22 wells is $(\overline{x}_1, \overline{x}_2, \overline{x}_3) = (46, -120, -466)$; subtracting from the x_i's, and finding the sums of the squares and products (the actual processing of the figures was carried out on the I.B.M.1620 Computer in the Memorial University of Newfoundland), we have the following variance-covariance matrix, S.

$$S = 1/21 \begin{pmatrix} 3136 & 1808 & 4207 \\ 1808 & 2251 & 6886 \\ 4207 & 6886 & 40432 \end{pmatrix}$$

Therefore the inverse = $S^{-1} = 21 \begin{pmatrix} .00062 & -.00063 & .00004 \\ -.00063 & .00157 & -.00020 \\ .00004 & -.00020 & .00005 \end{pmatrix}$.

Now, we should like the discriminant scores of the producing wells to be higher than those of the non-producers, so instead of the vector $(m_1 - m_2)$ we shall use the vector$(m_2 - m_1)$, and we shall assign a new well to population (1) of non-producers if its discriminant score is less than $\log_e(\$300{,}000)(.2)/(\$100{,}000)(.8) = -.2877$.

In this field study, $(m_2 - m_1) = (7.9,\ 13.5\ ,\ 57) = (d_1,\ d_2,\ d_3)$

Thus the coefficients in the discriminant function,$(b_1,\ b_2,\ b_3) =$

$(S^{11}d_1 + S^{12}d_2 + S^{13}d_3,\ S^{21}d_1 + S^{22}d_2 + S^{23}d_3,\ S^{31}d_1 + S^{32}d_2 + S^{33}d_3) =$

$(-.027,\ +.101,\ +.0085)$. thus the discriminant function,X, is:

$X = -.027x_1 + .101x_2 + .0085x_3$, and the mean discriminant score,$\overline{X}$, is:

$\overline{X} = -.027(46) + (.101)(-120) + (.0085)(-466) = -124 -12.12 -3.97 = -17.33$.

Now the discriminant score of the i'th well is $X_i - (-17.33)$, these scores are given in table 3.2, together with the part scores of the 3 variables contributing to the discriminant; this will yield 4 mappings which will help in the discovery of stratigraphic traps. We also perceive that on average, the variables contribute 7%,70% and 23% towards the discrimination, respectively.It appears therefore that the discrimination might be carried out using variable x_2 alone, and that , as a new exploratory hole is drilled, the depth at which we reach the BL zone is the key to whether we have a producer, that is if we reach the BL zone above -120 feet,subsea (0 being sea-level), we can be fairly happy about the prospect. In general, if one variable is mainly the cause of a well producing, we may discriminate simply by assigning a well to population (1) if the variable is below the mean for the field, and to population (2), if it is above it. Furthermore, if the variable is standardised, we can take into account the Bayesian a priori probabilities, and thus maximise the expected utility.

Referring back to theorem 3.2, we note that the costs of misclassification were brought in solely for economic reasons, they do not help in discriminating between wells that have already been drilled; they only help in decisions concerning a new well, or prospective well. Thus in testing the power of a discriminant, we assign a well to population (1) if its discriminant score X is less than $\log_e q_1/q_2$, that is in this case $\log_e 10/12 = -.1823$, hence the assignments in table 3.2.

TABLE.3.2:SPROULE FIELD,ALBERTA: DISCRIMINANT SCORES,PART SCORES,& POPULATIONS.

Discriminant Score.	Well No.		Population Actual	Population Assigned	Viking Score	BL Score	BSL QTZ Score
+.31	2	21	2	2	-.57	.60	.28
+1.29	6	21	2	2	-.51	1.50	.30
+1.12	7	21	2	2	-.32	1.10	.34
+1.40	8	21	2	2	-.35	1.50	.25
+ .87	11	21	2	2	-.16	.80	.23
+ .86	6	27	2	2	-.51	1.10	.27
- .07	11	27	2	2	-.08	-.10	.11
+1.48	12	27	2	2	+.60	.70	.18
+ .14	15	28	2	2	0.00	-.10	.24
+ .83	4	29	2	2	0.00	.60	.23

TABLE 3.2: Continued.

DISCRIMINANT SCORE	WELL NO.		POPULATION ACTUAL	POPULATION ASSIGNED	VIKING SCORE	BL SCORE	BSL QTZ SCORE
- .24	2	33	2	1	.19	-.50	.07
+ .22	4	33	2	2	.35	-.30	.17
-1.32	5	1	1	1	.24	-.90	-.66
-2.77	5	4	1	1	.41	-2.60	-.58
+ .64	1	22	1*	2	-.38	.00	.32
- .67	4	22	1	1	-.03	-.40	-.24
-3.05	13	24	1	1	.05	-1.60	-1.50
.23	4	27	1	2	0.00	.40	-.17
-.48	9	33	1	1	.41	-1.00	.11
-.67	10	33	1	1	.35	-.90	-.12
-.70	16	33	1	1	.10	-.80	-.01
-1.08	4	35	1	1	.05	-.60	-.53

*Is well number #1 22 really a producer; the BSL QTZ zone producing interval was not tested? Its score certainly indicates production. Omitting this well, the discriminant function is correct in 19 wells out of 21, that is over 90.5% successful. The part scores are useful both for mappings and for drilling decisions as exploratory drill-holes proceed. The following mappings further demonstrate the utility of the discriminant function in finding favourable locations to drill in the Sproule field, (see maps A_5, B_5, C_5, and D_5 on page 43), theoretically however the main aim is to test whether a new well is a non-producer belonging to population (1), or a producer to population (2); and the decision criterion is : assign a well $x = (x_1, x_2, x_3)$ to population (1) if $((x_1-46)(-.0013) + (x_2 + 120)($

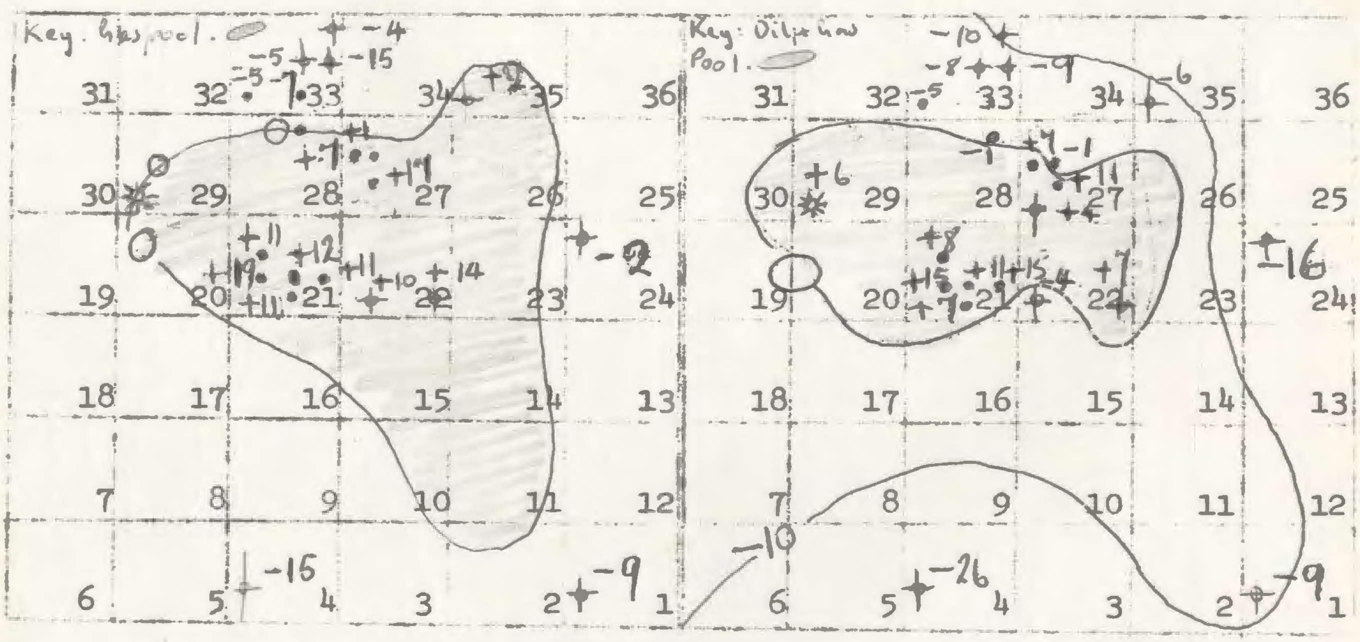
and to population (2) of producing wells if X is greater than or = -.29. We notice that this is a perfect example for applying the idea of discrimination on one variable alone, for if we choose variable x_3, the producing BSL QTZ sand, then the mean depth below sea-level is -466 feet; and subtracting this from each of the 22 wells yields the following discriminant scores:

a) POPULATION (1) NON_PRODUCERS: -69,-61,+27*,-26,-155,-19,+9,-14,-3,-44

b) POPULATION (2) PRODUCERS: +27,+29,+33,+24,+21,+25,+9,+16,+22,+21,+5,+16

*well no.#1.22,which may be a producer: interval not tested; so that if we discriminate by assigning a well with a negative score to population (1) and otherwise to population (2), then the discrimination is correct in 20/21 cases; that is over 95% accuracy.Furthermore this is a decision rule which can be applied on the drilling location.

FIELD STUDY NO.1: MAPPINGS.

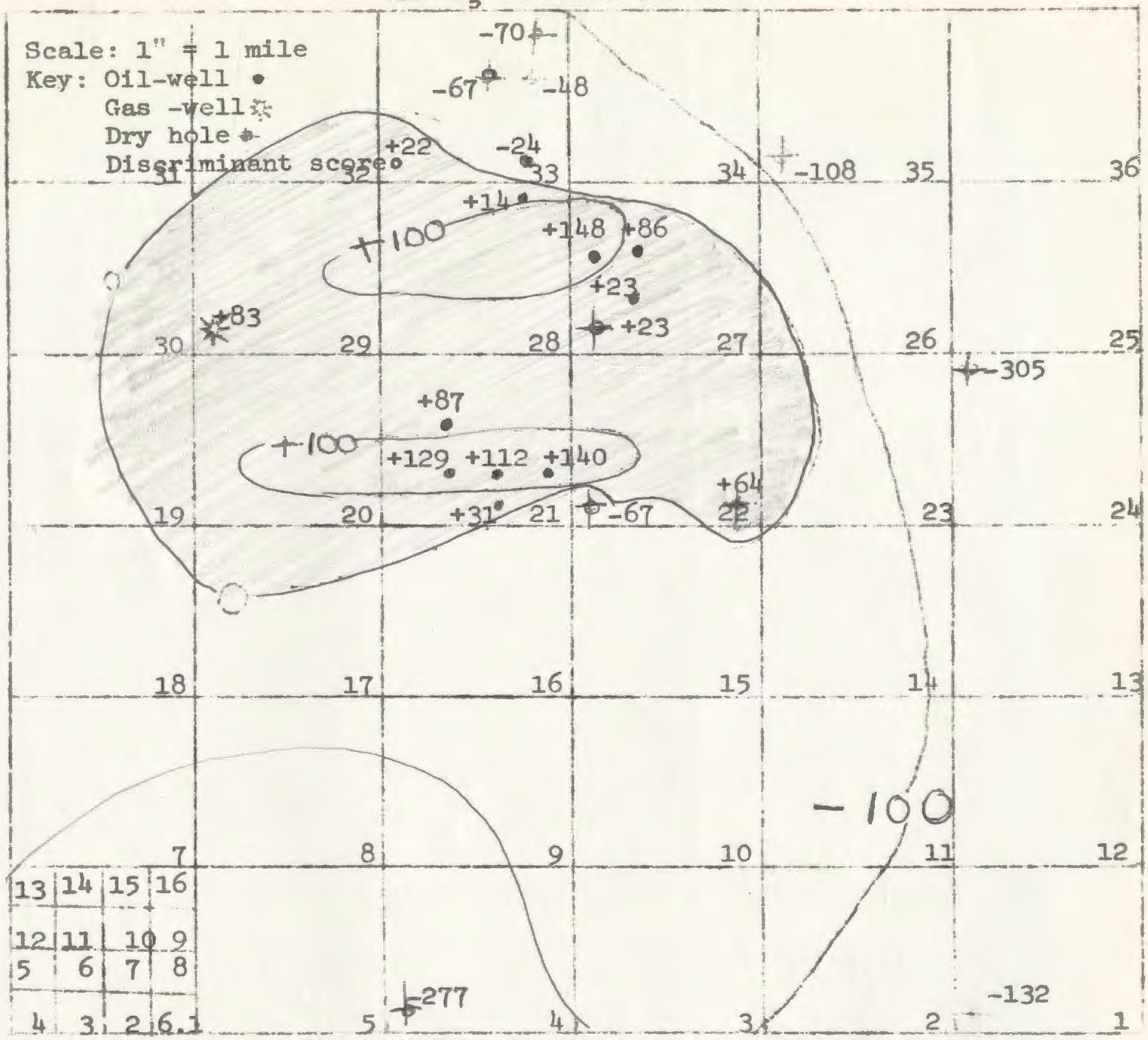


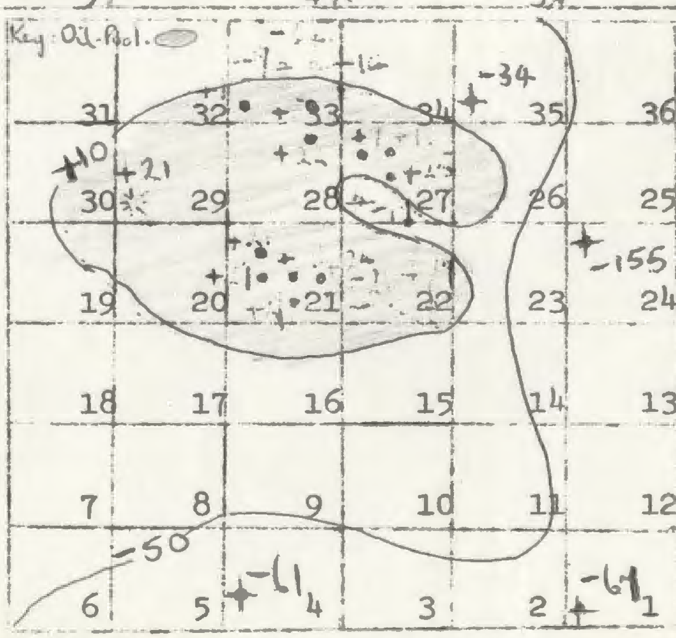
MAP B_5 VIKING. MAP C_5 BL. ZONE.

Scale:½" = 1 mile.

Key: Oil-well •
Gas well ✻
Discriminant score. e.g. +15.
Dry Hole

MAP A_5 DISCRIMINANT SCORE.





MAP D_5 BSL QTZ.

Each mapping indicates a similar region favourable to oil and gas production, but map A gives the strongest trends; so that here again the discriminant scores turn out to be the most useful exploration tool. The 0 contour marks the boundary of the producing field, and so the field may be far from drilled up.

Now the standard error of the discriminant score of individual wells is using theorem 3.1: $\sqrt{S_x}$ where $S_x = (\overline{X}_1 - \overline{X}_2) / (n_1 + n_2 - 2 - p + 1)$. Now the difference between the mean scores in this field is = $(-.027)(7.9) + (.101)(13.5) + (.0085)(57) = 1.698$, so that $S_x = 1.698 / (22 - 2 - 3 + 1) = .094$. Thus the standard error for individual wells is = .31.

The ratio of half the difference between the means to the standard error in individual wells is .849 / .31 = 2.74 . Assuming the discriminant scores are normally distributed, the probability of misclassifying a well is .0031, using normal distribution tables; thus in classifying new wells we may expect to assign nearly 99.7% to their correct populations.

In our second field study, we are given the following geological and geophysical data: x_1 is the thickness of the producing sand in feet; x_2 is the shaliness factor, m_s; and x_3 is the saturation ratio, s_{w_i} / s_w. The data for this study was collected by Howard Slack and Carl Otte from Oilfields in Texas and Oklahoma; and was given in their paper : "Electric log interpretations in exploring for Stratigraphic traps in shaly sands",published in the Bulletin of the American Association of petroleum Geologists, 1960. They define the shaliness factor, x_2, as $x_2 = m_r / s_w$ where m_r is the concentration of ions in the internal solution of the rock network in grams equivalent per litre, and s_w is the formation water saturation expressed as a fraction of the available pore-space. The saturation ratio, x_3, is defined as $x_3 = s_{w_i} / s_w$, s_{w_i} is the sand filtrate saturation of the invaded zone expressed as a fraction of the pore volume. Sample calculations of these factors are given in the appendix. The following data on which we shall formulate our discriminant function was given in the isopach and isopotential maps in their work:

TABLE 3,3: TEXAS AND OKLAHOMA OIL TRAPS IN SHALY SANDS.

Non-producers: Population (1).			Producing Wells : Population (2).		
x_{11} (feet)	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}
12	1.1	2.3	20	0.0	3.9
10	1.0	2.0	20	0.2	3.2
18	1.6	1.1	15	1.5	4.4
19	1.0	1.8	20	1.3	2.0
20*	0.1*	4.1*	17	0.8	4.8
14	3.4	2.4	20	0.6	4.5
8	4.6	2.2	19	0.9	3.1
9	4.7	1.0	20	0.2	3.2
17	0.8	0.1	16	0.2	2.9
10	0.8	3.8	12	0.8	2.0
14	0.3	2.2			

*This well was drilled in 1951, but the interval was not tested; the field was not discovered until 1955; had the 1951 hole been tested this would have been the discovery well; I have therefore ignored this well in formulating the discriminant function and substituted well 11 in its place. The above data was then processed on the I.B.M.1620 Computer, yielding the following results: $\overline{x}_1 = (13.1, 1.93, 1.89)$, $\overline{x}_2 = (17.9, .65, 3.4)$

$$S = 1/19 \begin{pmatrix} 17.1 & -3.38 & 1.32 \\ -3.38 & 1.85 & -.60 \\ 1.32 & -.60 & 1.52 \end{pmatrix} \quad \text{and } S^{-1} = 19 \begin{pmatrix} .091 & .162 & -.015 \\ .162 & .908 & .217 \\ -.015 & .217 & .757 \end{pmatrix}$$

Let $X = b_1x_1 + b_2x_2 + b_3x_3$ be Fisher's linear discriminant function, then

$b_1 = (.091)(4.80) + (.162)(-1.28) + (-.015)(1.51) = +.205$

Similarly $b_2 = -.057$ and $b_3 = +.79$.

Thus $X = .205x_1 + (-.057)x_2 + .79x_3$

Now the mean vector for the 20 wells $= (15.05, 1.29, 2.64)$

Thus $\overline{X} = .205(15.05) - .057(1.29) + .79\,(2.64) = 5.21$ is the discriminant score of the well mid-way between the means of the two populations.

Now, by theorem 3.4, we assign a new well $x=(x_1,x_2,x_3)$ to population (1) if its discriminant score, X, minus $\overline{X}$ is less than $\log_e c_2q_2/c_1q_1$, otherwise we assign the new well to population (2),; let the costs of misclassification be as before and let $q_1=.8$ and $q_2=.2$, then $\log_e c_2q_2/c_1q_1 = \log 3/4 = -.285$. The discriminant scores for the 20 wells and the results of the assignments are given in table 3.4. As before, we only take into account the costs in the case of a new well or prospective well, thus we have assigned a well to population one if x, its discriminant score is less than log 10/10 = 0 since there are 10 producing wells and 10 non-producers. As will be seen from table 3.4, the discrimination in this study is much more accurate since these are the data one would like to have on every well. It will be shown later that we can discriminate very well on the geophysical tests data alone, and of course the thickness of the producing sand is of obvious importance; hence the accuracy of the discriminant

TABLE 3.4: DISCRIMINANT SCORES AND POPULATION ASSIGNMENTS OF WELLS DRILLED IN SHALY SANDS.

	Discriminant Score.	Population Assigned
a) Non-producers:	-1.00	1
	-1.64	1
	- .74	1
	.05	2
	- .64	1
	-2.10	1
	-2.75	1
	-1.70	1
	- .21	1
	- .62	1
b) Producing Wells	+1.98	2
	+1.40	2
	+1.25	2
	+ .40	2
	+2.02	2
	+2.41	2
	+1.08	2
	+1.40	2
	+ .34	2
	-1.22	1

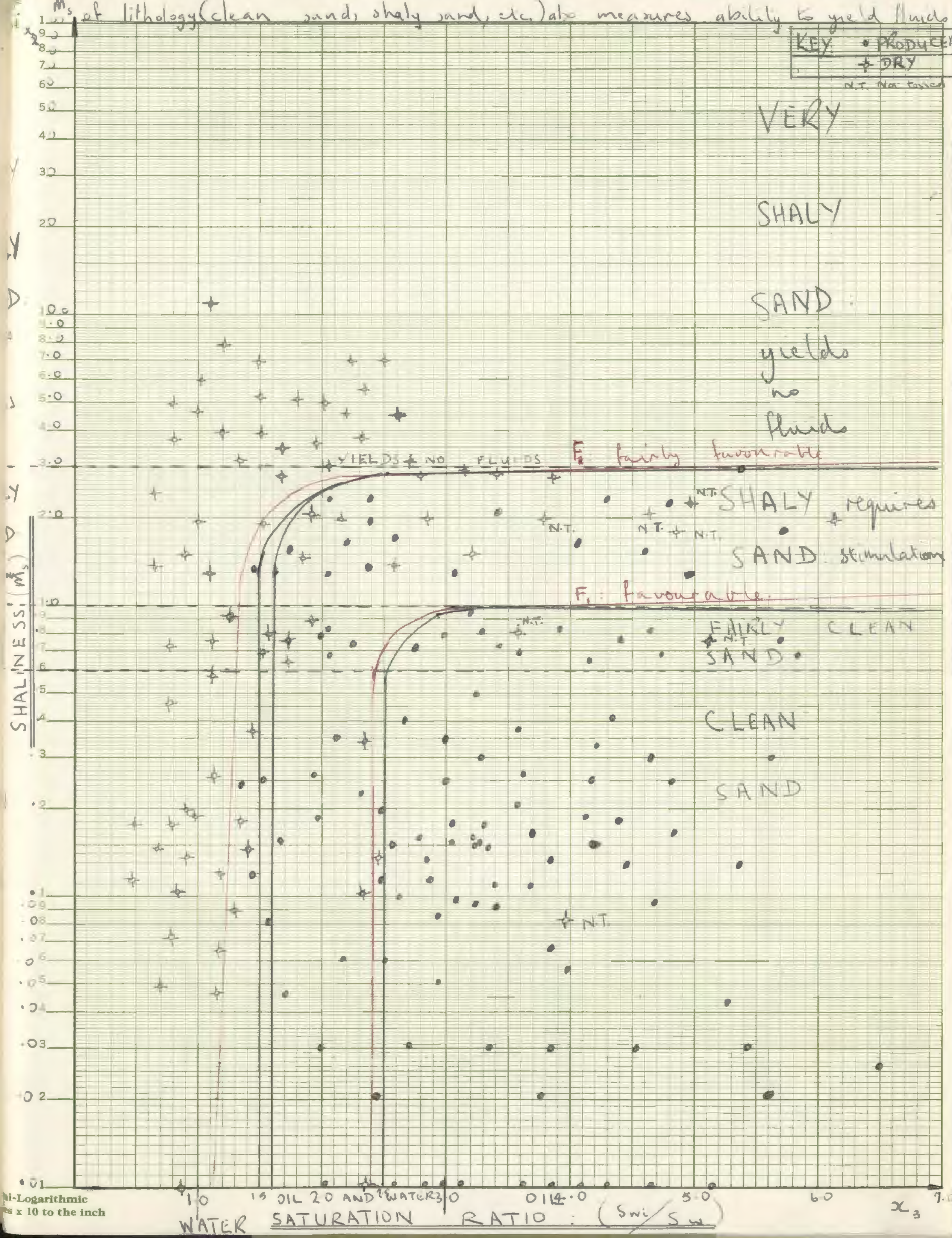
The discriminant function is accurate in 90% of the cases in this example. Slack and Otte gave the same data for 160 wells, in Oklahoma and Texas; in the appendix the discriminant score has been calculated for each well, and the ensuing population assignments are correct in 129 / 160 cases, (see appendix).

Using theorem 3.1, the standard error for individual wells is = $\sqrt{(6.34 - 4.07)} / 16 = .377$, and the ratio of half the difference between the means to the standard error = $\frac{1}{2}(6.34 - 4.07) / .377 = 3.01$ using slide-rule. Thus the probability of correctly classifying a new well using the discriminant function exceeds 99.8%.

Now referring to the petroleum-well performance graph, which illustrates the geophysical data of the 160 wells, we perceive that we may discriminate using these data alone, by dividing the space into 3 regions whose boundaries are formed by fitting rectangula hyperbolae by eye. We then have regions which are (1)Favourable , (2) Fairly favourable and (3) Unfavourable to petroleum production.

We therefore define a favourability factor,F, as follows: $F = (x_3 - 1.5)(3 - x_2)$; and we notice that $(x_3 - 1.5)(3 - x_2) \geq .15$ defines a region in which the vast majority are producers, thus if $F = .15$ or greater we can use this as our criterion for drilling a well, or for assigning tt to population (2) of producing wells. Furthermore if we define $F^* = (x_3 - 2.5)(1 - x_2)$ we note that $(x_3 - 2.5)(1 - x_2) \geq .15$ defines a region in which almost every well is a producer. Thus $F^* \geq .15$ defines a region of near certainty;

PETROLEUM WELL PERFORMANCE: Shaliness is a measure of lithology (clean sand, shaly sand, etc.) also measures ability to yield fluids



functions F and F* both being derived by fitting rectangular hyperbolae empirically. Now F* is the function F with origin translated from (3,1.5) to (1,2.5); we can therefore define the 3 regions by using F alone since the favourable region is reached when F = (2.5-1.5)(3-1) + .15 = 2.15. Thus discrimination using F alone is achieved by defining the following regions: (1) When $F < .15$: Unfavourable for petroleum production

(2) When $.15 \leq F < 2.15$: Fairly favourable

(3) When $F \geq 2.15$: Favourable for petroleum production

where $F = (x_3 - 1.5)(3.0 - x_2)$ as previously defined, and where at least one of the factors in the product are positive. If both $x_3-1.5$ and $3.0-x_2$ are negative the well belongs to the other branch of the curve defining a region of unfavourable prospects. Thus if both factors are negative, assign the well to population (1) of non-producers; without calculating the value of F.The favourability scores for the 20 wells are:

a)POPULATION (1) : 1.52,1.0,-.56,.6,7.44*,-.36,-1.12,both negative,-3.08, 5.06,1.89.

b)POPULATION (2) : 7.2,4.76,4.35,.85,7.26,7.2,3.36,4.76,3.92,1.10.

Thus the discovery well* would easily have been found using this method, and furthermore when 3 regions are defined we have 20/21 correctly assigned, and even the incorrect one was technically a producer, it was non-commercial only because the producing sand was too thin.

When the faverability scores for all the 160 wells are calculated, 148/160 were correctly assigned; and this is particularly significant as the wells came from oilfields in both Texas and Oklohoma.

The second field study was continued by calculating the discriminant function, using all the 160 wells to try to sharpen the discrimination. The results were: $\overline{x}_1 = (13.12, 2.08, 1.94)$; $\overline{x}_2 = (16.35, .48, 3.28)$, so that

$(d_1,d_2,d_3) = (3.23,-1.60,1.34)$;

$S = \begin{pmatrix} 22.69 & -1.55 & .66 \\ -1.55 & 1.57 & -.61 \\ .66 & -.61 & 1.82 \end{pmatrix}$ so that $S^{-1} = \begin{pmatrix} .047 & .045 & -.001 \\ .045 & .777 & .243 \\ -.001 & .243 & .641 \end{pmatrix}$

Hence $(b_1,b_2,b_3) = (.081,-.771,.456)$ using the slide rule.

$\therefore$ The discriminant function for the 160 control points is :X, where $X = .081x_1 - .771x_2 + .456x_3$ and hence we assign a well to population (2) if $X - 1.66 \geq -.133$ where 1.66 is the discriminant score midway between the discriminant scores of the mean non-producer and the mean producing well. The results of this discrimination are given in the appendix,where the discriminant scores are given for all the 160 wells and 129/160 are correctly classified,almost 81% accuracy.

Since x_2, and x_3 are not altogether independent, a further study is now carried out using thickness of bed,x_1, and favourability factor F to see if this gives a sharper discriminant ; the results are:

$(\bar{x}_1, \bar{x}_2) = ((13.12,.93),(16.35,4.73))$.

$S = \begin{pmatrix} 22.69 & 2.18 \\ 2.18 & 12.82 \end{pmatrix}$ so that $S^{-1} = \begin{pmatrix} .044 & -.008 \\ -.008 & .080 \end{pmatrix}$

Hence, $(b_1, b_2) = (.112, .278)$ and $X = .112x_1 + .278\,F$ is the discriminant function, and, since $\bar{X} = 2.44$ is the discriminant score midway between the two populations and $-.133$ is $\log_e$ of the ratio of the a priori probabilities , we assign a well to population (2) if its discriminant score X, is such that $X-2.44 \geq -.133$.

The one hundred and sixty discriminant scores are given in the appendix, and 147/160 are correctly assigned if we define 3 regions as in the favourability factor example; ie. Unfavourable $X < 1.7$,Fairly favourable $1.7 \leq X < 2.7$

and favourable: $X \geq 2.7$.

A second discriminant rule may be defined by using the principal component scores defined in Chapter 2 .Let m_1 be the mean component score of the non-producers and m_2 the mean principal component score of the producing wells. Then, we may assigna new well to population (1),if its principal component score is less than$(m_1 + m_2)/2$, and otherwise assign it to population (2). Unfortunately, Krumbein did not specify which wells produced in his paper, so that we are unable to test the power of the discriminant.

A third discriminant function we shall consider is given by Kendall and Stuart in "The Advanced Theory of Statistics"; and it is applicable only when the correlations between the x_i's are equal.We shall consider it however since this is the situation in John C. Griffiths' work in "Computers in Mineral Industries: a Symposium",Stanford University, California entitled "A Statistical Approach to the study of Potential Oil Reservoir Sandstones".WE shall give an account of this paper after we have formulated the discriminant function applicable.

When the correlations are all equal to r, then the latent roots of the correlation matrix,

$$R = \begin{array}{c c} & \begin{array}{cccc} x_1 & x_2 & \cdots & x_p \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ x \\ \vdots \\ \cdot \\ x_p \end{array} & \begin{pmatrix} 1 & r & \ldots & r \\ r & 1 & r & r \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ r & r & \ldots & 1 \end{pmatrix} \end{array}$$

may be shown to be:

$$b_1 = 1 + (p-1)r,$$

$$b_2 = b_3 = \ldots = b_p = (1-r)$$

The variation therefore contains a main component, $y = 1/\sqrt{p} \sum_{i=1}^{p} x_i$, corresponding to b_1, and we take the size component,Q,proportional to this, $Q = \sum_{i=1}^{p} x_i = \sqrt{p}\, y.$; and the variance of $Q = pb_1 = p(1+(p-1)r)$.

No other component is outstanding so we define a shape component,P, as $P = \sum_{i=1}^{p}(w_i - \bar{w})x_i \,/\, \bar{w}$, where $w_i = \bar{x}_{1_i} - \bar{x}_{2_i}$ for each i.

Then the variance of $P = \sum_{i=1}^{p} ((w_i - \bar{w})/\bar{w})^2 (1-r)$ and the covariance of $(P,Q) = (1+(p-1)r)\sum_i (w_i - \bar{w})/\bar{w} = 0$. Thus the discriminant X is of the form $X = aQ + P$ such that it maximizes $(\bar{x}_1 - \bar{x}_2)^2 / \text{Var.x}$.

Writinq $D_p = P_1 - P_2$ and $D_q = Q_1 - Q_2$, we have then to maximise:

$$(aD_q - D_p)^2 / (a^2 \text{var } Q + 2a \,.\text{cov}(P,Q) + \text{var } P)$$

Differentiating partially with respect to a, we have:

$$2(aD_q + D_p)D_q(a^2\text{var}Q + 2a \text{ cov}(P,Q) + \text{var } P) - (2a \text{ var } Q + 2\text{cov}(P,Q))(aD_q + D_p)^2 = 0$$

Hence $a = (D_q \text{var } P) / (D_p \text{var } Q)$

Hence $a = (1-r) / (1 + (p-1)r)$

and the discriminant $X = (1-r)Q / (1 + (p-1)r) + P$.

We use the size and shape variate to discriminate as in the general method.

FIELD STUDY NO. 3: In our 3rd field study we use the data given in John Griffiths' paper previously referred to , in which he gives data on the quartz grains contained in the Maxton Sandstone,(Mississippian),West Virginia. Griffiths gives the following correlation matrix concerning the petrographic properties of the sandstone; the 3 variables being: x_1 = matrix, x_2 = quartz length, and x_3 = Grain breadth.

PETROGRAPHIC PROPERTIES OF MAXTON SANDSTONE

CORRELATION MATRIX, AND DISCRIMINANT.

$$R = \begin{array}{c c} & \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} & \begin{pmatrix} 1 & .8 & .8 \\ .8 & 1 & .9 \\ .8 & .9 & 1 \end{pmatrix} \end{array}$$

Since the correlations are almost equal, we take r=.8, and p=3

Thus $X = .2\, Q / (1 + 2(.2)) + P$,

$X = Q/7 + P$ is the discriminant function. Thus if we had the petrographic data on sands from samples of producing wells and non-producers, we could use this function to assign a new well to its correct population.

Although the discriminant functions are applicable for general use, and for use in provinces, basins, and petroleum fields with few control points (wells), a sharper discriminator can be calculated for each oil-field as soon as control data becomes available. The results derived in this chapter demonstrate the accuracy of the discriminant function and its two main uses as a decision function:(1) To assign a well to its correct population as down-hole (stratigraphic and geophysical) information becomes available, and (2) to predict the locations favourable to production by plotting the discriminant scores. The last function should be very useful in predicting whether a sandstone is likely to be barren or whether it is likely to produce, especially if it is found from outcrops of the formation. One might be able to predict more accurately even without control points.

When the exploratory hole has been drilled, or is being drilled, the discriminant function is decisive in predicting whether the well will produce or not; simply by calculating the discriminant score or part score.

CONCLUSION

Decision functions are valuable tools for use in exploration and development of petroleum wells, fields, basins and provinces. The general principals laid down in the thesis are sound, but the functions may be sharpened in particular locations by calculations based on data from local control points (wells).

The discriminatory functions of chapter 3 are all new, and the idea of discriminating on the spot by merely using the point midway between the means of the non-producing and producing populations, using the high points of the producing sand alone is a revelation. Another valuable tool, enabling the petroleum operator to discriminate without the use of a computer when the saturation factor and shaliness factor are known, is the use of a new favourability factor, F; a location being favourable if the value of $F = (s_{w_i} / s_w - 1.5)(3 - m_s)$ exceeds or = .15 , provided the factors in the multiplication are positive. Specimen calculations of the shaliness measure and favourability criterion are given in the appendix.

Despite the increase in the use of the computer in the petroleum industry during the past decade, optimal use of well information is still far from being achieved. This paper has been written to demonstrate how any measurable, relevant information can be used for more economic exploration and development of petroleum resources. The data one discriminates upon in a particular location can of course only be decided upon by team of petroleum experts on the spot. I have attempted to define how the information should be processed, when this decision has been taken.

When the exploratory hole is drilled, the discriminant function is decisive in predicting whether the well will produce or not.

BIBLIOGRAPHY.

Computers in the Mineral Industries:Third Annual Conference,1963;
Stanford University, California

APPLICATIONS of Statistics,Operations Research, and Computers in the Mineral Industries: International Symposium,Colorado School of Mines,1964.

Americam Association of Petroleum Geologists Bulletins,1959 - July,1967.

"Statistical Decision and Related Techniques in Oil and Gas Exploration": Gordon M. Kaufman: Prentice-Hall,1963.

"Geology of Petroleum": Leversen; Freeman.

"Petroleum Production": Jones; Reinholdt.

"Introduction to Multivariate Statistical Analysis":Anderson;Wiley.

"A Course in Multivariate Statistical Analysis":Kendall;Griffin.

"The Advanced Theory of Statistics":Kendall & Stuart: Griffin.

"Concepts and Methods of Experimental Statistics":Fryer;Allyn & Bacon.

"Statistical Inference":Keeping; Van Nostrand.

"Multivariate Statistical Methods":Morrison; McGraw-Hill.

"World Oil",April,1967 and other copies 1960-67.

"Decisions under Uncertainty; Drilling Decisions by Oil & Gas Operators": Grayson (Harvard,1960)

"Contributions to Mathematical Statistics": R.A.Fisher;Wiley.

"Discriminant Analysis": Tatsmoka and Tiedeman; Review of Educational Research,Vol. XXIV,No.5.,Dec.,1954.

APPENDIX.

SPECIMEN CALCULATIONS: Shaliness, Saturation Ratio, and Favourability Factors. Part of an electric log of s shaly sand drill-hole in Oklahoma is sketched below. The producing interval is 4200-4217 ft. Data required for the calculation are as follows:

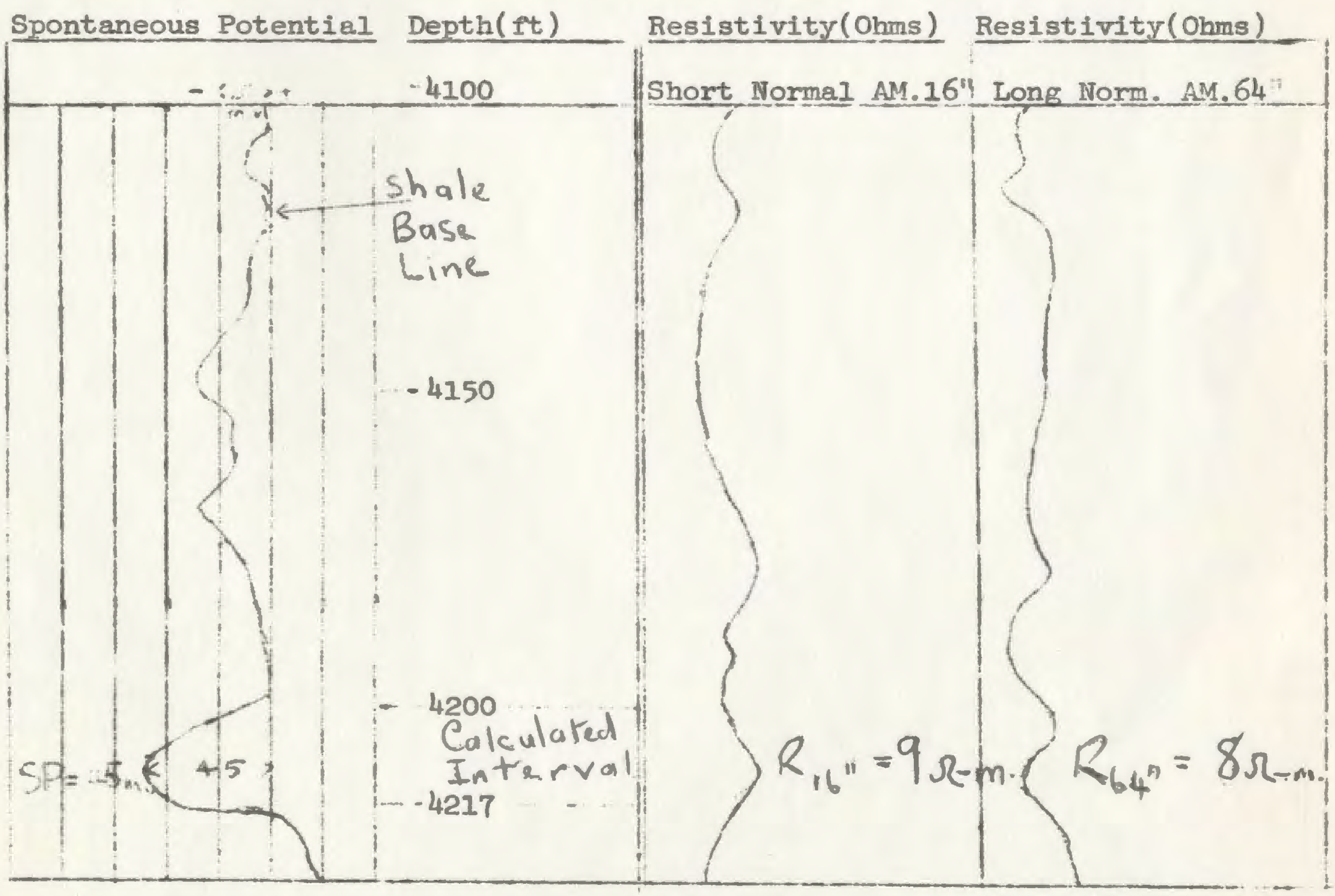
Diameter = $8^3/4''$, formation temperature = 100° F = t

Mud Resistivity = 1.2 ohm-metres; spontaneous potential, SP_s = -45 m.v.

Resistivity 16" Normal, $R_{16''}$ = 9 ohm metres, and $R_{64''}$ = 8 ohm-metres.

Resistivity of water at 100° F = R_w = .04 ohm-metres.

PART OF ELECTRIC LOG OF SHALY SANDS IN OKLAHOMA.



1. Shaliness factor Calculation: $m_s = m_r/s_w$, $SP_s = -k \log \dfrac{(m_s+2.15m_w)}{m_s + 2.15m_{mp}}$ ..(1)

From the figure, SP_s = -45 millivolts.

k_t = .21 T where T is the formation temperature in degrees absolute (Kelvin) = ((100-32)/1.8 +273)(.21)=65.3

Now M_w= p.p.m. Na.11/58x1000 = 175000/58000 = 3 mol. wts/litre

R_{mf} = .95 ohm-metres and M_{mf}=4500/58000 = .078 mol. wts/litre.

Substituting in equation (1), antilog 45/65.3 = $\dfrac{m_s + (2.15)(3)}{m_s + 2.15(.078)}$

Therefore m_s=shaliness factor=1.4 gms equiv./litre.

2. Calculation of Saturation Ratio s_{wi}/s_w.

$SP_s = -k_t \log (R_i/R_t \times S_{w_i}/S_w \times (m_s+2.15(s_{w_i}/s_w))/(m_s + 2.15m_{mf})$ ..(2)

Now $R_{16''}/R_m$= 9/1.2=7.5 and $R_{64''}/R_m$=8/1.2=6.7, R_i/R_m=8, R_t/R_m=6.0.

Therefore R_i/R_t=8/6=1.33, and substituting in (2) gives:

antilog 45/65.3=1.3 s_{wi}/s_w (1.4+2.15S_{wi}/S_w x m_{mf})/(1.4 +2.15 x.078

4.9 = 1.3 S_{wi}/S_w (1.4 +.168S_{wi}/S_w)(/)(1.4+.168)

Hence .218 $(S_{wi}/S_w)^2$+1.82(S_{wi}/S_w) - 7.68 = 0

sothat the saturation ratio S_{wi}/S_w = 3.1

3.Calculation of Favourability Criterion,F.

$F= (S_{wi}/S_w - 1.5)(3 - m_s)$

so that, in this example, F= (3.1 - 1.5)(3 - 1.4) =2.56,

indicating since it is greater than 2.15 that this is a producing well.