Mathematical Geology, Vol. 22, No. 3, 1990

Bayesian Estimation of Undiscovered Pool Sizes Using the Discovery Record[1]

Lawrence D. Stone[2]

This paper presents a Bayesian version of the classical model of Kaufman et al. for the order of discovery of hydrocarbon pools and the distribution of their sizes in a geologically homogeneous area. Using the model, a Bayesian method is developed for estimating the distribution of the size of the undiscovered pools using the information from the discovery record. This method avoids most of the arbitrary choices required by the modified maximum likelihood method developed by Lee and Wang. As an example, this method is applied to the same Bashaw reef data on which Lee and Wang demonstrated their approach. For this case, the Bayesian approach produces sharly lower estimates of undiscovered resources.

KEY WORDS: petroleum resources, pool size distribution, Monte Carlo.

INTRODUCTION

Kaufman et al. (1975) present a model for the distribution of the hydrocarbon pool sizes in a geologically homogeneous area and postulate a probabilistic model for the discovery sequence of pools from this area. Briefly, Kaufman et al. assume that there is some number N of pools in the area. The sizes of these N pools have a distribution equal to that of N independent draws from a log-normal distribution with parameters μ and σ^2, that is, the distribution with density function f defined by

$$f(y) = (y\sigma\sqrt{2\pi})^{-1} \exp\left(-(\ln(y) - \mu)^2/2\sigma^2\right) \qquad \text{for } y > 0$$

The parameters μ and σ^2 are unknown.

The discovery sequence is assumed to follow a probabilistic model that depends on an unknown parameter β, where $-\infty < \beta < \infty$. Suppose that $S = (S_1, \ldots, S_J)$ is the vector of undiscovered pool sizes. Typically, pool sizes

[1]Received 31 March 1988; accepted 15 September 1989.
[2]Metron Inc., 1481 Chain Bridge Road, McLean, Virginia 22101.

0882-8121/90/0400-0309$06.00/1 © 1990 International Association for Mathematical Geology

are measured in terms of the volume they occupy. The probability that pool i is the next to be discovered is given by

$$\frac{S_i^\beta}{\sum_{j=i}^{J} S_j^\beta}$$

Suppose that $X = (X_1, \ldots, X_N)$ is the vector of the N pool sizes in the play. Let i_j be the index of the j^{th} pool discovered, and let $I_j = \{i_i, i_2, \ldots, i_j\}$ for $j = 1, \ldots, N$ with I_0 indicating the null set. The probability that pools $i_1, i_2, \ldots, i_n$ will be the first n discovered and in that order is

$$\prod_{j=1}^{n} \frac{X_{ij}^\beta}{\sum_{i \notin I_{j-1}} X_i^\beta}$$

Lee and Wang (1983a,b, 1985) present a modified maximum likelihood method of estimating the parameters N, β, μ, and σ given a discovery sequence. They proceed by choosing a set of values for N (e.g., N_1, N_2, $\ldots$ N_K). For each value N_k, they find the maximum likelihood estimates and confidence intervals for μ, and σ. Using these intervals they select a set of values of μ and σ to associate with N_k. They then compute the 25^{th} and 75^{th} percentile values for the pool sizes for each combination of N_k, μ, and σ. This procedure is repeated for each N_k. The user then picks a choice or choices of parameter values, which yield a set of pool size statistics that "matches" the sizes of the discovered pools. Lee and Wang (1985) give an example of such a match in Fig. 1, taken from that reference.

Having decided upon a match, the user assigns ranks (e.g., first largest, second largest pool in the play) to the discovered pools as shown (Fig. 1). The parameters chosen for the match are $N = 80$, $\mu = 6.0$, $\sigma^2 = 3.0$. Having specified N and the ranks of the discovered pools, the distribution of the size of the undiscovered pools is independent of β. Thus β drops out of Lee and Wang's analysis at this point. The ranks not assigned to discovered pools become estimates of the rank and size of undiscovered pools. For example (Fig. 1), the three largest discovered pools are assigned ranks 1–3. The fourth and fifth are assigned ranks 5 and 6. This means that the fourth largest pool in the play is estimated to be undiscovered. The resulting uncertainty intervals for the size of undiscovered pools (Fig. 2) is taken from Lee and Wang (1985).

There are a number of difficulties with this approach. The first is that the user has to make many arbitrary choices in order to obtain an estimate of the size and number of undiscovered pools. For example, the user has to choose values of N, μ, and σ, and assign ranks to the discovered pools. As Lee and Wang point out, the assignment of ranks to discovered pools can be rather arbitrary. For example, it would be reasonable to assign the 4^{th} and 5^{th} largest

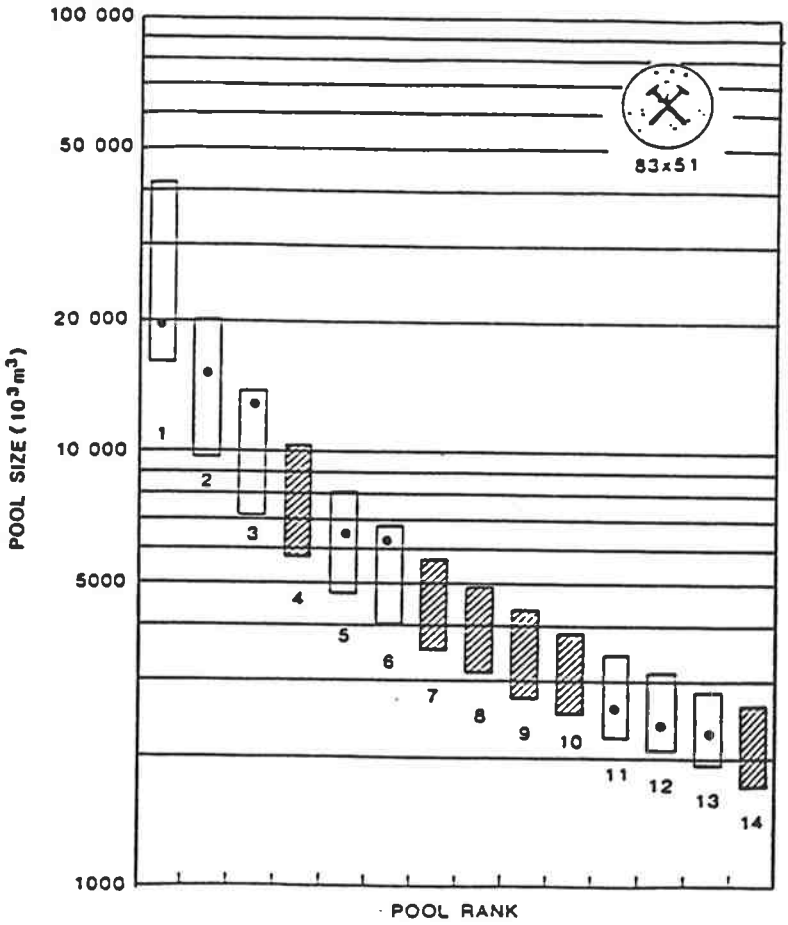

Fig. 1. Individual pool sizes by rank for $N = 80$, the Bashaw Reef play. The pool sizes are expressed by 25th and 75th upper percentiles. Dots indicate pool reserves: shaded areas are predicted undiscovered pool sizes. Note: $\mu = 6.0$, $\sigma^2 = 3.0$.

discovered pools rank 4 and 5, respectively (Fig. 1). This would result in the 6th largest pool in the play being estimated as the largest undiscovered pool rather than the 4th as shown (Fig. 1).

In general, the requirement that the user be forced to choose specific values for parameters that may be uncertain is undesirable and should be avoided.

A second difficulty occurs in the maximum likelihood estimate for N. For example, the Bashaw reef data in Lee and Wang (1985) produces a maximum likelihood estimate of 38 for the value of N which is equal to the number of discovered pools. Thus, the maximum likelihood estimate is that no undiscov-

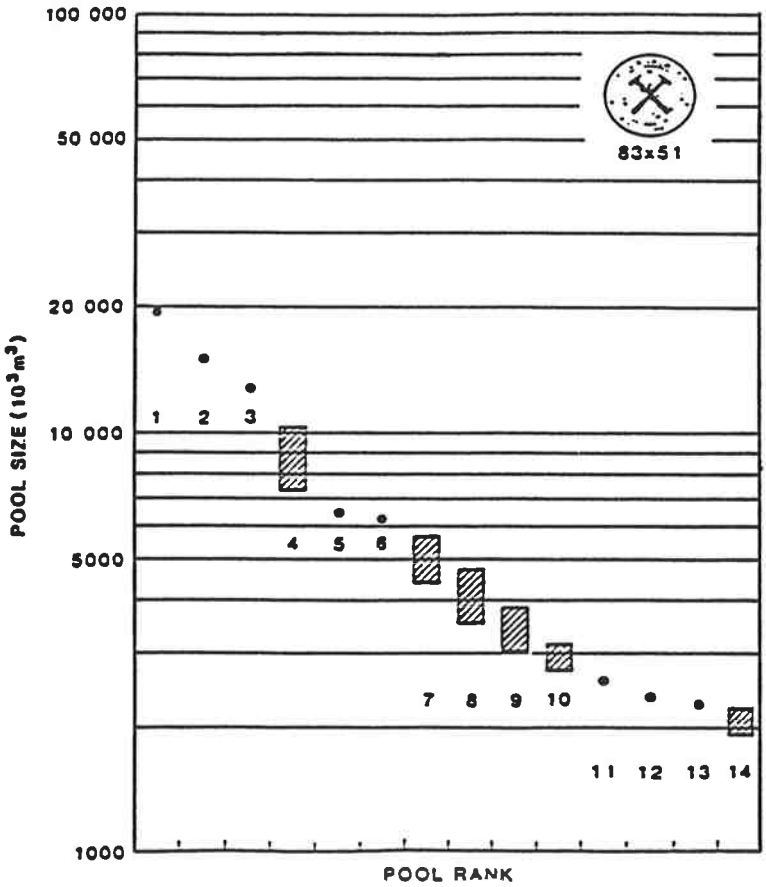

Fig. 2. Individual pool sizes by rank constrained to the discovery record and $N = 80$, the Bashaw Reef play. Symbols as for Fig. 1. Note: According to the *Daily Oil Bulletin* (Calgary, June 9, 1983) it appeared that the 7th pool had been discovered; $\mu = 6.0$, $\sigma^2 = 3.0$.

ered pools exist. This forced Lee and Wang to make a rather arbitrary choice for the value of N, and the values of μ and σ^2 consistent with N.

A third difficulty is that the resulting estimates of the size and number of undiscovered pools take no account of the geometry of the play, especially the size and locations of the discovered pools in the play region or of the dry holes in the region.

Grace (1986) discusses some of the advantages and difficulties in using Lee and Wang's approach.

This paper proposes a Bayesian method for estimating the size of undiscovered pools conditioned on a discovery sequence. This method overcomes the first and second difficulties discussed above. To overcome the third difficulty, a model of the discovery process is needed that, among other things, accounts for the geometry and dry holes in the play region.

The first section describes the Bayesian model that is used. The basic method used to compute distributions conditioned on the discovery sequence is derived and a Monte Carlo implementation of this method is described. The final section gives examples computed using this method applied to the Bashaw reef data.

BAYESIAN MODEL

A Bayesian version of the model given in Kaufman et al. (1975) is proposed. Specifically, the unknown parameters N, β, μ and σ^2 are assumed to have a prior distribution on their values. This prior may be obtained from geological and seismic information, or from plays that are similar to the one under consideration. Let $\theta = (N, \beta, \mu, \sigma^2)$ denote a vector consisting of these four parameters, and let Θ be the space of possible values of θ. Let g be the probability density function for the prior on θ. Note, the term probability density function is used even when discrete distributions may be involved. Given a value of θ, this Bayesian model reduces to the one used by Kaufman et al. (1975) and Lee and Wang (1985). For convenience, the notation $f(\cdot \mid \theta)$ will be used to refer to the log-normal density function obtained by using the values of μ and σ^2 in the vector $\theta = (N, \beta, \mu, \sigma^2)$.

THE BASIC CONDITIONAL EXPECTATION

Recall that $X = (X_1, \ldots, X_N)$ denotes the pool size vector and that I_j denotes the ordered set of indices of the first j pools discovered for $j = 1, \ldots, N$. Given θ, the pool sizes are distributed as independent draws from a log-normal distribution with density $f(\cdot \mid \theta)$ and the probability density for the pool size vector X is

$$\prod_{j=1}^{N} f(X_j \mid \theta)$$

The probability that the first n pools discovered have the ordered set of indices I_n given X and θ is

$$\prod_{j=1}^{n} \frac{X_{i_j}^{\beta}}{\sum_{i \notin I_{j-1}} X_i^{\beta}}$$

The joint density for X, I_n, and θ is

$$g(\theta) \prod_{j=1}^{N} f(X_j|\theta) \prod_{j=1}^{n} \frac{X_{ij}^{\beta}}{\sum_{i \notin I_{j-1}} X_i^{\beta}} \tag{1}$$

Observe that if any ordered subset I_n' of n elements of $\{1, 2, \ldots, N\}$ is chosen and the joint density for X, I_n' and θ is computed, an expression identical to (1) is obtained but with the indices rearranged.

When a discovery sequence is observed, only the size of the pools is known and not their indices in the set of pools. To emphasize this fact, denote the discovery sequence by $(Y_1, \ldots Y_n)$ to indicate that Y_1 could correspond to any X_i for $i = 1, \ldots N$, etc. There are $(N)_n \equiv N(N-1) \cdots (N-n+1)$ ordered subsets of n elements chosen from a set of N elements. For each choice, the above density function is obtained. Thus, the density function for the discovery sequence $Y = (Y_1, \ldots, Y_n)$, the pool size vector, X, and θ can be written as

$$(N)_n\, g(\theta) \prod_{j=1}^{n} f(Y_j|\theta) \prod_{j=n+1}^{N} f(X_j|\theta) \prod_{j=1}^{n} \frac{Y_j^{\beta}}{\sum_{k=j}^{n} Y_k^{\beta} + \sum_{k=n+1}^{N} X_k^{\beta}} \tag{1'}$$

This expression is obtained by evaluating expression (1) for $I_n = \{1, 2, \ldots, n\}$ and $X_i = Y_i$ for $i = 1, \ldots, n$, and then multiplying by the factor $(N)_n$ to account for the other $(N)_n - 1$ ordered choices from $\{1, 2, \ldots, N\}$.

At this point, it is convenient to split the field size vector X into two components, $X = (Y, S)$, where $S = (S_{n+1}, \ldots, S_N)$ is the size vector for the undiscovered fields. Observe that knowing X and Y is equivalent to knowing Y and S. Replace X_j by S_j in expression (1') for $j = n+1, \ldots, N$ and define

$$\varphi(Y, S, \theta) = (N)_n\, g(\theta) \prod_{j=1}^{n} \frac{f(Y_j|\theta) Y_j^{\beta}}{\sum_{k=j}^{n} Y_k^{\beta} + \sum_{k=n+1}^{N} S_k^{\beta}} \prod_{j=n+1}^{N} f(S_j|\theta)$$

Then φ is the joint density for the discovery sequence Y, the vector S of undiscovered field sizes, and θ. Let $\varphi^*(S, \theta \,|\, Y)$ be the conditional density for S and θ given Y. Then

$$\varphi^*(S, \theta \,|\, Y) = \frac{\varphi(Y, S, \theta)}{\int \varphi(Y, S, \theta)\, dS\, d\theta} \tag{2}$$

Define

$$h_1(S, \theta, Y) = (N)_n \prod_{j=1}^{n} \frac{f(Y_j|\theta)\, Y_j^{\beta}}{\sum_{k=j}^{n} Y_k^{\beta} + \sum_{k=n+1}^{N} S_k^{\beta}}$$

$$h_2(S, \theta, Y) = g(\theta) \prod_{j=n+1}^{N} f(S_j|\theta)$$

Then

$$\varphi^*(S, \theta|Y) = h_1(S, \theta, Y)\, h_2(S, \theta, Y)/C$$

where C equals the denominator on the right-hand side of Eq. (2).

Let ψ be any real valued function defined on $R^{N-n} \times \Theta$. Let $Z = (S, \theta)$. It is desired to evaluate $E[\psi(Z)|Y]$ by Monte Carlo methods where Z has the density φ^*. To do this, make I independent draws $Z_1, \ldots, Z_I$ from the distribution with density h_2. (Note, h_2 is a probability density.) Observe that

$$\frac{1}{I} \sum_{i=1}^{I} h_1(Z_i)\, \psi(Z_i) \to E_2\left[h_1(Z)\, \psi(Z)\right] \text{ as } I \to \infty \tag{3}$$

where E_2 indicates expectation with respect to the distribution defined by h_2. Observe that

$$E_2\left[h_1(Z)\, \psi(Z)\right] = \int_{R^{N-n}\times\Theta} h_1(S, \theta, Y)\psi(S, \theta)h_2(S, \theta, Y)dS\, d\theta$$

$$= C \int_{R^{N-n}\times\Theta} \varphi^*(S, \theta|Y)\, \psi(S, \theta)dS\, d\theta \tag{4}$$

Combining Eqs. (3) and (4), it follows that

$$(IC)^{-1} \sum_{i=1}^{I} h_1(Z_i)\, \psi(Z_i) \approx E[\psi(S, \theta)|Y] \tag{5}$$

Equation (5) is the basic result that will be used to obtain conditional distributions on $Z = (S, \theta)$ given the discovery sequence Y.

THE MONTE CARLO APPROACH

From Eq. (5) it is clear how to produce Monte Carlo draws that will provide a means to compute any expectation or distribution on (S, θ) conditioned by Y. Proceed by drawing I independent values $(S(i), \theta(i))$ for the vector $(S,$

θ). These are drawn from the distribution with density h_2 by first drawing θ_i from the density g. This yields

$$\theta(i) = (N_i, \beta_i, \mu_i, \sigma_i^2)$$

Then make $N_i - n$ independent draws from $f(\cdot \mid \theta(i))$ to obtain $S(i) \equiv (S_{n+1}(i), \ldots, S_{N_i}^i(i))$. Finally, calculate

$$W_i = h_i\big(S(i), \theta(i), Y\big)$$

The result of the i^{th} replication in this Monte Carlo is the vector

$$\big(W_i, N_i, \beta_i, \mu_i, \sigma_i^2, S_{n+1}(i), \ldots, S_{N_i}(i)\big) \qquad \text{for } i = 1, \ldots I$$

For later computational convenience, store the S_k values in descending order i.e., $S_{n+1}(i) \geq S_{n+2}(i) \geq \ldots \geq S_{N_i}(i)$. Having performed all I replications, compute

$$C' = \sum_{i=1}^{I} W_i / I \tag{6}$$

By setting $\psi(Z) \equiv 1$ in Eq. (5), the reader can see that C' is an approximation to C.

CALCULATING CONDITIONAL DISTRIBUTIONS AND EXPECTATIONS

Having performed the above Monte Carlo simulation and stored the results, any distribution or expectation conditioned on Y can be calculated. The following are some examples.

Number of Pools

The expected total number of pools, N, given Y, is

$$E[N \mid Y] \approx (C'\, I)^{-1} \sum_{i=1}^{I} W_i N_i \tag{7}$$

and the distribution on the number of pools is computed by

$$Pr\{N \leq K\} \approx \sum_{\{i:N_i \leq K\}} W_i / (C'I) \tag{8}$$

Size of the k^{th} Largest Undiscovered Pool

Similarly, the expected value and distribution of the size of the k^{th} largest undiscovered pool is given by

$$E\big[S_{n+k} \mid Y \text{ and } N \geq n + k\big] \approx \big(D(k)\big)^{-1} \sum_{\{i:N_i \geq n+k\}} W_i S_{n+k}(i) \tag{9}$$

and

$$Pr\{\text{size of the } k^{\text{th}} \text{ largest undiscovered pool} \le K\} = \sum_{\{i:S_{n+k(i)} \le K\}} W_i/(D(k)) \tag{10}$$

where

$$D(k) = \sum_{\{i:N_i \ge n+k\}} W_i$$

Estimate of β

The discovery rate parameter is estimated in a similar fashion. In the standard Bayesian fashion, the expectation of the posterior on β is used as the estimator for β, that is,

$$\beta^* = E[\beta \mid Y] = (C'I)^{-1} \sum_{i=1}^{I} W_i \beta_i$$

Estimates of μ and σ^2

Posterior estimates of μ and σ^2 may be obtained in an analogous fashion. Specifically,

$$\mu^* = E[\mu \mid Y] = (C'I)^{-1} \sum_{i=1}^{I} W_i \mu_i$$

$$\sigma^{2*} = E[\sigma^2 \mid Y] = (C'I)^{-1} \sum_{i=1}^{I} W_i \sigma_i^2$$

EXAMPLE

This section presents estimates made using the methods described above and the Bashaw Reef discovery data from Lee and Wang (1985).

Description of Example

This example repeats Lee and Wang's analysis but uses the Bayesian methods described in this paper. The discovery sequence listed in Lee and Wang (1985) (Table 1) is used, and the following prior distributions on β, μ, σ, and N are assumed:

- β, the discovery parameter, has a uniform distribution on [0, 1].
- μ, the log-normal mean parameter, is distributed normally with mean 6.0 and variance 1.0.

Table 1. Discovery Sequence for the Bashaw Reef Play

Rank	Pool name	Discovery data[a]	Pool size (10^3 m^3)
5	Stettler A	1949-03-30	6150
8	Duhamel B	1950-08-04	2240
18	Fenn Big Valley A	1950-08-10	509
25	New Norway	1950-08-12	318
10	Bashaw A	1951-05-27	1600
23	Stettler South	1951-11-28	394
11	Malmo A	1951-12-07	1510
20	Nevis Devonian	1952	429
19	Malmo D	1952-02-07	480
38	Fenn Big Valley G	1952-02-21	48
24	Fenn Big Valley E	1952-05-01	329
36	Fenn Big Valley C	1952-06-11	110
26	Stettler B	1952-08-29	300
3	Clive	1952-09-18	13100
4	Erskine	1952-11-06	6390
17	Ewing Lake	1952-12-03	516
14	West Drumheller	1953-01-08	1250
29	Fenn Big Valley B	1954-06-14	261
6	Fenn Big Valley F	1954-10-13	2870
2	Wimborne	1956-02-08	15000
32	Duhamel A	1956-08-02	191
1	Innisfail	1957-04-22	19700
27	Wood River	1957-06-17	294
15	Chigwell B	1959-01-26	631
12	Buffalo A	1960-12-30	1410
7	Lone Pine Creek	1962-12-08	2350
21	Chigwell A	1964-01-22	427
28	Bashaw B	1965-07-04	264
37	Malmo C	1965-07-21	71
31	Nevis C	1967-10-28	222
16	Buffalo Lake B	1967-11-11	556
9	Haynes	1968-06-27	1670
30	Nevis B	1968-11-08	238
34	Penhold	1968	183
33	Nevis D	1969-03-11	191
13	Nevis E	1970-08-30	1270
22	Nevis F	1970-10-27	400
35	Mikwan	1970-12-09	131

[a]Discovery date is assigned as the spud date of the discovery well. Dates are not available for the Nevis Devonian and the Penhold pool. They were assigned by Lee and Wang (1985).

- σ^2, the log-normal variance parameter, is gamma distributed with mean 3.0 and variance 1.5 (standard deviation 1.22).
- N, the total number of pools in the play is uniformly distributed over the integers in the interval [40, 100].

In addition, assume that these distributions are independent, so that

$$g(\theta) = g_1(N)\, g_2(\beta)\, g_3(\mu)\, g_4(\sigma^2)$$

where

g_1 is a uniform discrete distribution over [40,100]

g_2 is a uniform density function on [0, 1]

g_3 is a normal density function with mean 6 and variance 1

g_4 is a gamma density function with mean 3 and variance 1.5

The prior distributions on μ and σ^2 are chosen to correspond to Lee and Wang's assumptions. In particular, Lee and Wang's results are based on assuming $\mu = 6.0$ and $\sigma^2 = 3.0$. Thus, the mean of the priors is taken to be 6.0 and 3.0, respectively. For β, a uniform distribution on [0, 1] is used. For N a uniform distribution on [40, 100] is used because this covers almost all values of N considered by Lee and Wang (1985) in the maximum likelihood table (Table 2) of that reference. By contrast Lee and Wang base their estimates of undiscovered pool sizes on assuming $N = 80$.

Using the methods described above, a Monte Carlo sample of 64,000 points was produced and used to make estimates of the posterior distributions of the following quantities conditioned on the discovery sequence:

β = the discovery parameter

N = the total number of pools in the play

μ = the log-normal "mean" parameter

σ^2 = the log-normal "variance" parameter

50% intervals for the size of the k^{th} largest undiscovered pool

90% intervals for the size of the k^{th} largest undiscovered pool

Density function for the 1st, 2nd, and 3rd largest undiscovered pool

Density function for the size of an undiscovered pool

Distribution of the size of the total undiscovered resource

Conclusions

First a word of caution about these conclusions. They are based on prior distributions developed by the author to show an example of the application of the Bayesian approach. The priors are taken to correspond to Lee and Wang's analysis, so that the results in this paper can be compared to those in Lee and Wang's. The estimates produced for this example should *not* be interpreted as this author's best estimates for the undiscovered resources in the Bashaw Reef play. A serious effort to produce such estimates would include obtaining the latest data from the play and discussion with experts knowledgeable about the play to determine the priors to be used for the estimation.

The results shown below lead to the following conclusions:

1. The posterior estimate of β is 0.32, indicating a weak relationship between the size of a pool and the order of its discovery. If the pools were spherical in shape and the probability of discovery were proportional to their projected area onto the surface of the earth, $\beta = 2/3$ would be expected.
2. The posterior distribution on N shows that larger values of N are less likely than smaller ones, but the trend is a gentle one. This means that the discovery sequence for the Bashaw Reef yields little information about the total number of pools in the play. This is expected to be true whenever the posterior estimate of β is small.
3. The rather arbitrary choice of $N = 80$ by Lee and Wang (1985b) leads to large estimates of the size of the k^{th} largest undiscovered pool and the total undiscovered play potential compared to those obtained from a uniform [40,100] prior on N.

Discussion of Results

The computations described above produced histograms of the posterior distributions for β, N, μ, and σ^2 along with the estimated mean and standard deviation of these distributions (Figs. 3–6). The 50% and 90% containment intervals for the posterior distribution of the k^{th} largest undiscovered pools were calculated for $k = 1, \ldots, 20$ (Figs. 7 and 8). The procedure produced empirical density functions for the size of the 1^{st}, 2^{nd}, and 3^{rd} largest undiscovered pool (Figs. 9–11) as well as the density function for the size of an (unordered) undiscovered pool (Fig. 12). The total amount of undiscovered resource in the play was also estimated (Fig. 13).

The system developed to produce these estimates consists of two programs that run on an Apollo DN3000 or DN4000 workstation (restricted versions of these programs have been developed for IBM AT compatible computers). The first program generates Monte Carlo vectors in the manner described in the

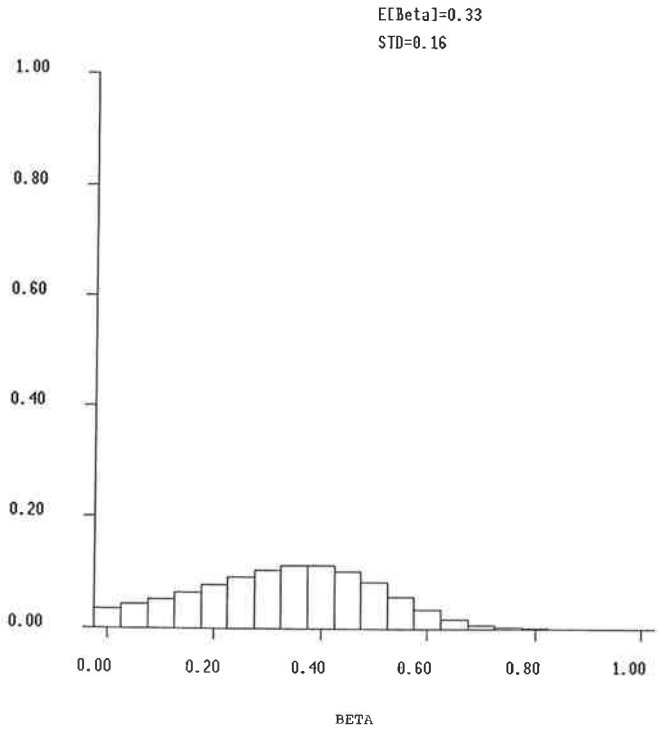

Fig. 3. Posterior on β.

section entitled, Monte Carlo Approach. Because the weights for the points are unequal and can vary by several orders of magnitude, the system throws out vectors with very small weights and continues to generate points until it retains the desired number. The resulting vectors are stored in a file for use by the second program. The second program allows the user to interactively produce the statistical estimates described above (Figs. 3–12) as well as others not shown (e.g., the distribution of the size of the 10^{th} largest undiscoverd pool). The user has the option of choosing the range of values and number of bins on the horizontal axis for the histogram and density estimates. Figures shown here are direct copies of figures that appear on the screen. Ranges and the number of bins were chosen to provide a balance between smoothness and resolution. Running on a DN4000, the program that generated the Monte Carlo vector file required 6 hrs to produce 64,000 retained vectors. After an initial period to read in the replication file, the interactive statistical analysis program produced most of the graphs shown here with little delay (e.g., 5–10 sec).

The posterior distribution on β (Fig. 3) is consistent with the maximum likelihood values for β as a function of N, given (Table 2) by Lee and Wang

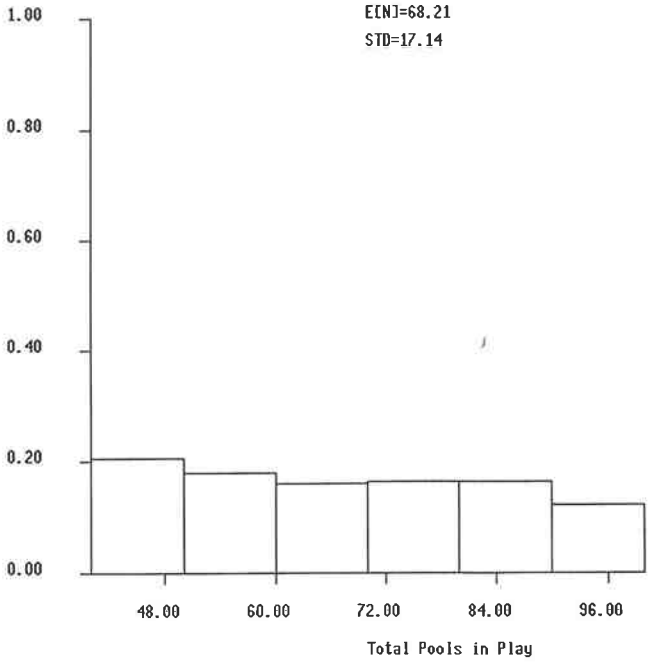

Fig. 4. Posterior on N.

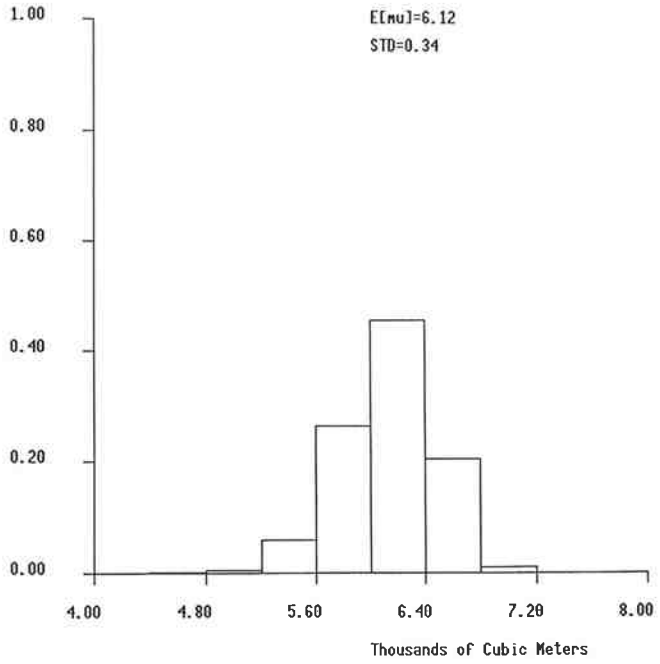

Fig. 5. Posterior on μ.

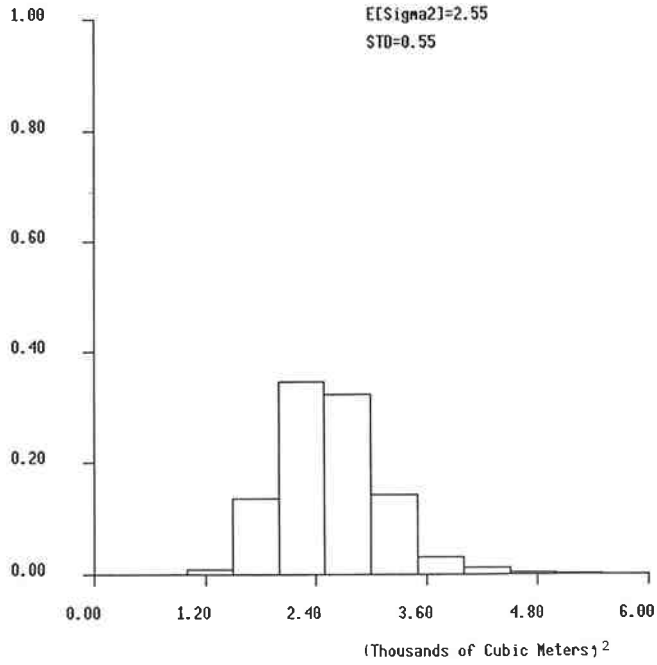

Fig. 6. Posterior on σ^2.

(1985). $E[\beta] = 0.33$ is the expectation of the posterior distribution on β. This is the Bayes estimator for β. The value of 0.33 indicates that there is not a strong relationship between the size of a pool and the probability of discovering it.

The posterior on N (Fig. 4) shows a mildly decreasing trend in likelihood as the value of N increases. This results from the weak relationship between pool size and discovery order. In this case, the discovery sequence contains little information about the number of pools in the play.

The posterior on μ (Fig. 5) has almost the same mean as the prior but a reduced standard deviation. This indicates the data are consistent with $\mu = 6.0$. The mean of the posterior on σ^2 (Fig. 6) moved down to 2.55 from the prior mean of 3.0 and standard deviation decreased, indicating that 2.55 is a better estimate of σ^2 than 3.0.

Both the 50% and 90% containment intervals for the size of the 20 largest undiscovered pools (Figs. 7 and 8) are lower than the corresponding intervals (Fig. 2) in Lee and Wang (1985). There are two reasons for this. First, Lee and Wang rather arbitrarily chose $N = 80$. Because this is in the high end of the interval [40,100], this tends to increase estimates of undiscovered pool size over those obtained from the Bayesian analysis, which assumes that the prior

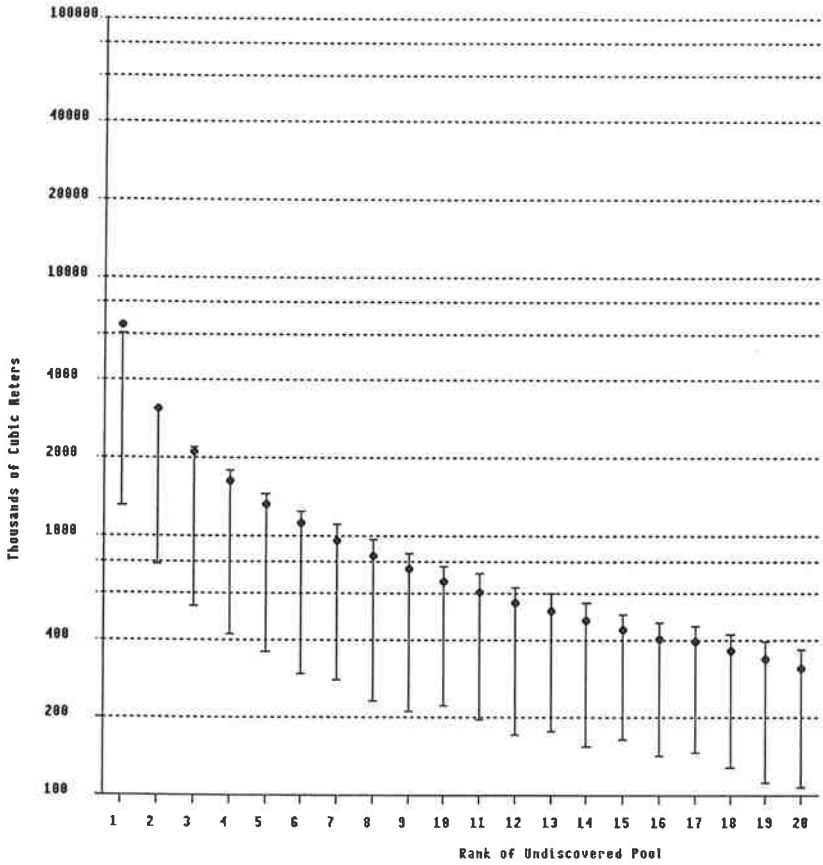

Fig. 7. 50% intervals for the k^{th} largest undiscovered pool. Note: the dots indicate the mean values of the size of the k^{th} largest undiscovered pool. The 50% interval covers [0.25, 0.75].

on N is uniform over [40,100]. Second, Lee and Wang's methodology forces one to assign ranks to discovered pools. Lee and Wang (Fig. 1) chose an assignment which assumes that the fourth largest among all the pools in the play is still undiscovered. One could just as reasonably have assigned the 5 largest discovered pools to ranks 1–5, leaving the 6^{th} largest among all pools undiscovered. This would substantially reduce Lee and Wang's estimates of the size of the largest undiscovered pool. The methodology used in the Bayesian approach effectively averages over all possible rank assignments, weighted by their likelihood, to produce estimates of the size of the k^{th} largest undiscovered pool. This avoids the problem of having to make an arbitrary choice between two almost equally reasonable choices.

In order to obtain estimates of the parameters of the log-normal densities that approximate the posterior densities on the size of the three largest undiscovered pools (Figs. 9–11), the logarithms of the sample pool sizes were taken and the mean (μ) and standard deviation (SD) of these logarithms were computed.

The posterior density for the size of an undiscovered pool (Fig. 12) is a reasonable fit to a log-normal one. Note that μ for the undiscovered pools is reduced to 5.38 compared to 6.12 for the distribution of all pool sizes, discovered plus undiscovered (Fig. 5).

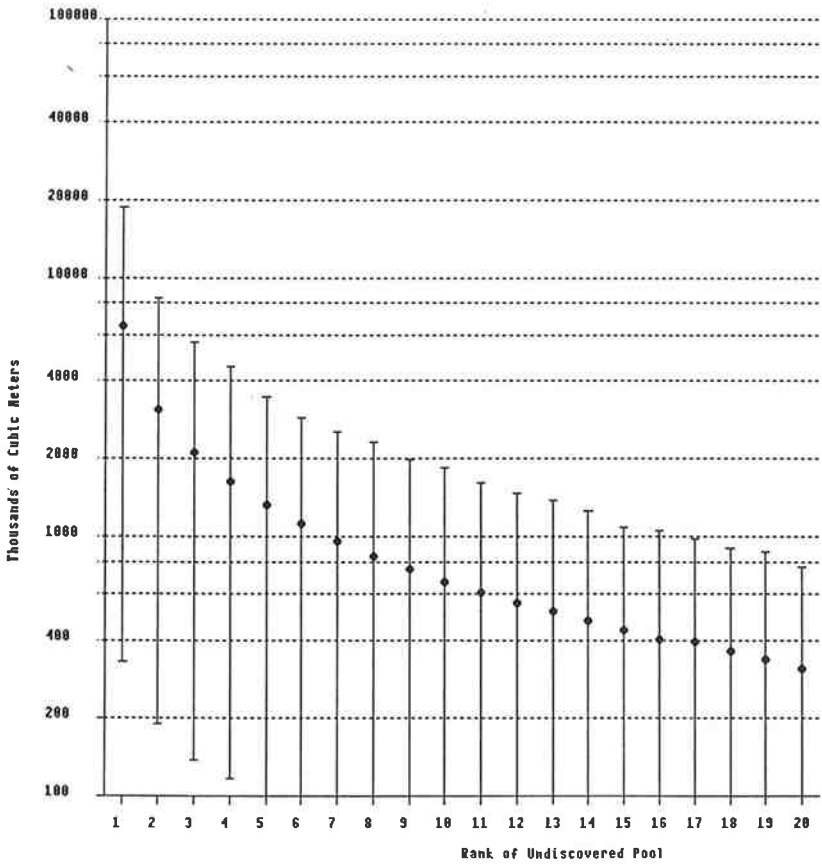

Fig. 8. 90% intervals for the size of the k^{th} largest undiscovered pool. Note: The dots indicate the mean values of the size of the k^{th} largest undiscovered pool. The 90% interval covers [.05, 0.95].

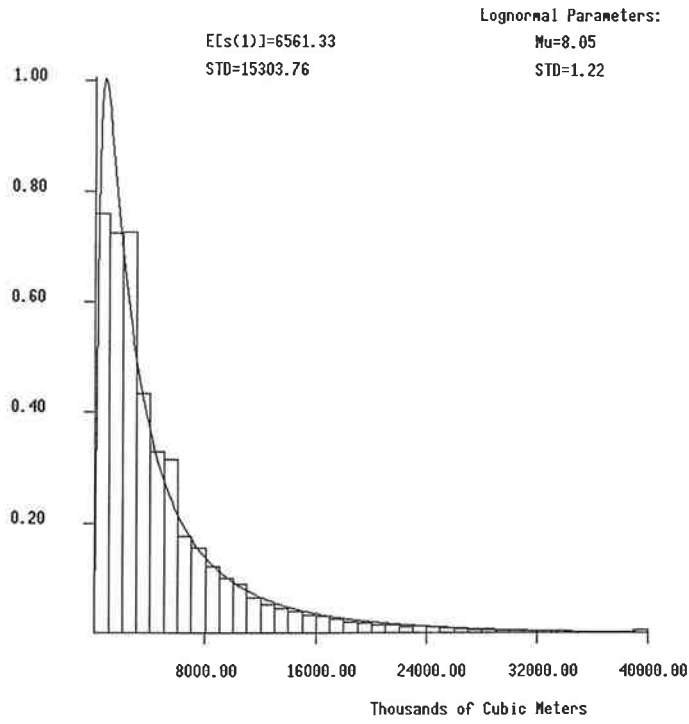

Fig. 9. Posterior density for the size of the largest undiscovered pool.

The probability distribution for the total resource remaining undiscovered is shown (Fig. 13) with bars that represent bins 8 million cubic meters in size. The height of each bar is the probability of the total undiscovered resource falling in that bin. The figure is not designed to show these probabilities accurately, but the data from which the plot is made yields the following. Let T be the total undiscovered resource measured in millions of cubic meters. Then

$$Pr\{T \leq 8\} = 32\%$$
$$Pr\{T \leq 32\} = 77\%$$
$$Pr\{32 \leq T \leq 56\} = 14\%$$
$$Pr\{T \leq 56\} = 91\%$$

The above analysis shows that there is a 14% probability of the total undiscovered resource being between 32 and 56 million m^3. This is in comparison to Lee and Wang's estimate of a 90% chance of the total undiscovered resource being between 32 and 50 million m^3. Clearly, the Bayesian method yields sharply reduced estimates of undiscovered resources in this case.

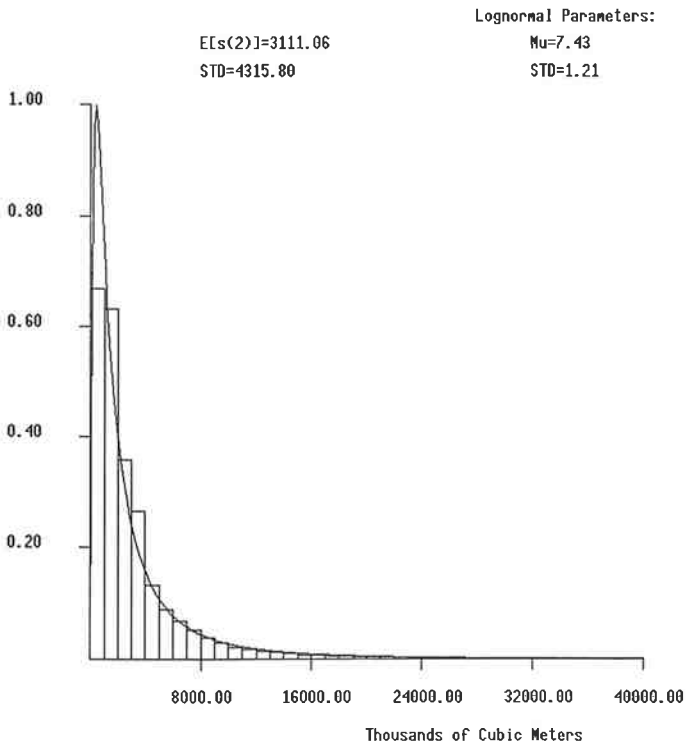

Fig. 10. Posterior density for the size of the 2nd largest undiscovered pool.

DISCUSSION AND CONCLUSIONS

A Bayesian model has been developed for estimating undiscovered pool sizes, using the discovery record from a play, and applied to the same Bashaw Reef data analyzed by Lee and Wang, using a modified maximum likelihood method. At this point it would be natural for the reader to ask himself two questions.

Is the Bayesian approach really an improvement over the Lee and Wang approach?

How would this Bayesian methodology be applied to plays other than the Bashaw Reef?

Is the Bayesian Approach Better?

The answer to this question is *yes*. Any good method of estimating undiscovered pool sizes from the discovery record requires subjective inputs. In the case of Lee and Wang's approach, the subjective inputs enter in several places.

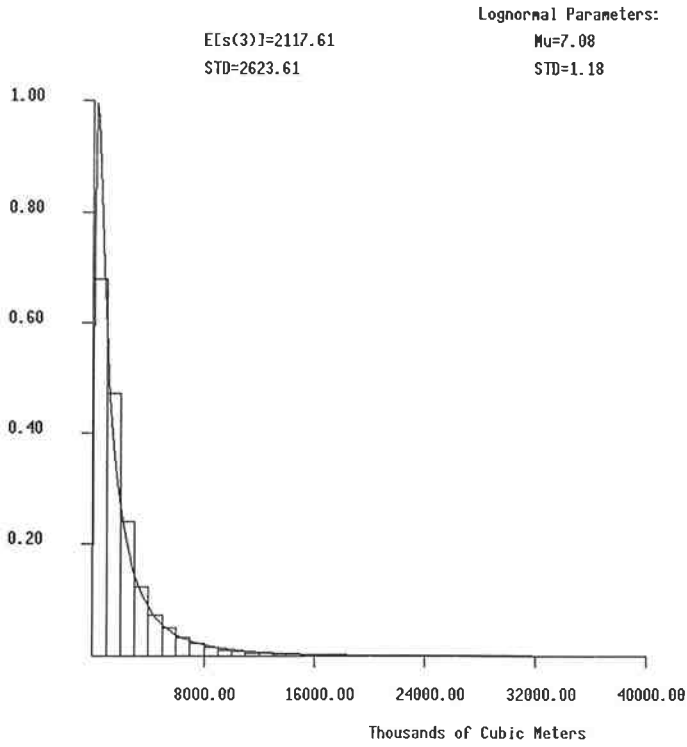

Fig. 11. Posterior density for the size of the 3rd largest undiscovered pool.

Lee and Wang's method is not a pure maximum likelihood method. The pure maximum likelihood estimate of the total number of pools in the play is 38, which is exactly the number of pools discovered (at of the time of the analysis). This estimate would imply that no pools are left to be discovered, which is not a satisfactory conclusion. This led Lee and Wang to establish a subjective matching criterion that resulted in the choice of $N = 80$ for the total number of pools in the play, as well as $\mu = 6.0$ and $\sigma^2 = 3.0$ for the parameters of the log-normal distribution for the pool sizes in the play. Additional subject inputs were made through the assignment of ranks (within the totality of pools, both discovered and undiscovered) to the already discovered pools.

The Bayesian approach requires subjective inputs for the distributions on N, μ, and σ, the same parameters that Lee and Wang determined by their subjective matching procedure.

The Bayesian approach also requires subjective inputs for the distribution of β. Lee and Wang obviate the need for β by specifying, in a subjective manner, the rank of the discovered pools. Thus, both methods require the analyst

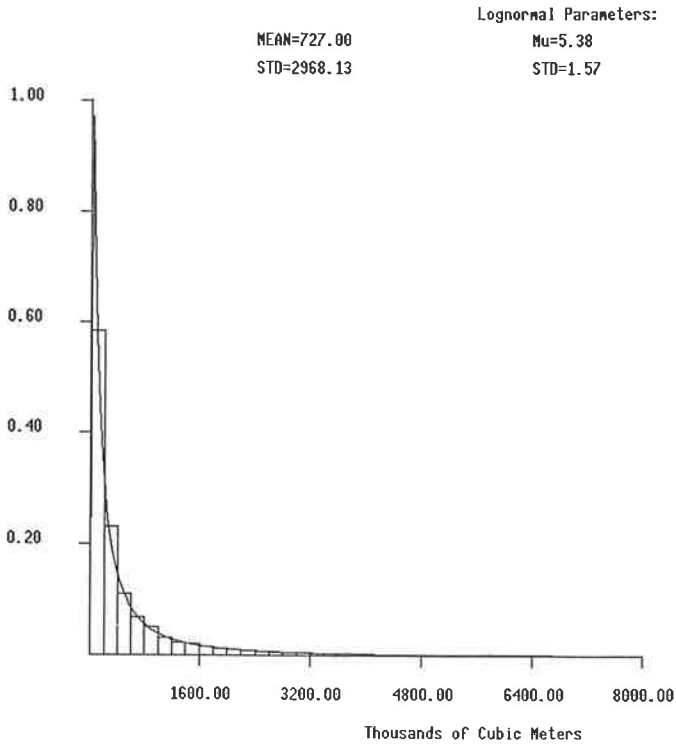

Fig. 12. Posterior density for the size of an undiscovered pool.

to subjectively specify virtually the same parameters. The author contends that any reasonable method based on the Kaufman et al. model must in some fashion specify N, μ, σ, and β in order to perform its analysis. These parameters are necessary to define the problem.

Given the necessity to specify these parameters or their equivalents, the author feels that the Bayesian approach is superior for four reasons. First, it allows the analyst to specify his estimates of the required parameters *and his degree of uncertainty about the estimates*. Second, these subjective estimates are identified clearly in the analysis. Third, the estimates are used in a probabilistically and mathematically consistent methodology to produce estimates of undiscovered resources. For example, the Bayesian methodology correctly averages over all possible rank assignments for the observed pool sizes, weighting them by their likelihood. In contrast, Lee and Wang's estimate is based on only one of the possible rank assignments. Fourth, a greater variety of parameters can be estimated (e.g., the posterior on N).

An implementation question arises in the examples presented in this paper.

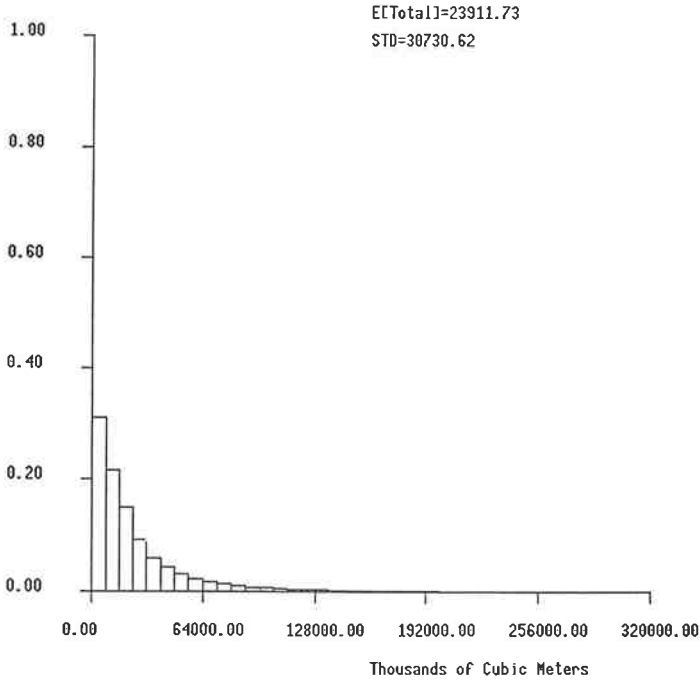

Fig. 13. Probability distribution for total undiscovered resource.

The numerical method that is used to calculate the estimates of undiscovered pool sizes and other parameters involves a Monte Carlo simulation of the posterior distributions. This naturally raises questions about the statistical error in the estimates. The unequal sample weights, which are crucial for efficient generation of the posterior distributions, make it difficult to answer these questions. On the other hand, it is difficult to obtain confidence intervals from standard maximum likelihood estimators, much less the subjectively modified ones used by Lee and Wang. This appears to be an unsolved problem for both approaches. In the case of the Monte Carlo approach, this problem can be handled in an operationally reasonable fashion by incrementally increasing the sample size used until the estimates settle down. For the example calculated here, this required about 6 hrs of computer time on a modestly priced computer workstation. For a large company with more capable computers, even much larger problems should present no difficulty.

Application of the Bayesian Approach to Other Plays

The example presented in this paper deals with the Bashaw Reef play because it allows the use of publicly available data and provides a direct compar-

ison of the Bayesian approach to the modified maximum likelihood approach of Lee and Wang. How would one apply the Bayesian approach to other plays? Of course, the key question here is how to obtain the subjective estimates of N, μ, σ, and β? The answer is by consulting geologists and explorationists familiar with the play in question and others similar to it. One must elicit subjective probability distributions for these parameters from these people. Methods for eliciting such probabilities are described in Savage (1971), Lindley et al. (1979), Spetzler and Stael Von Holstein (1975), Morris (1977), and Winkler (1981). These methods have worked in developing Bayesian approaches to finding submarines, people lost at sea, and sunken treasure (see Stone, 1983). One of the crucial features of this approach is that the uncertainty in the estimate is specified in addition to the estimate itself. Thus, for a play where a lot of exploration has taken place, the estimates will tend to be based on substantial amounts of data and will have small uncertainties. In relatively unexplored plays, the estimates will tend to have large uncertainties. This is the way the estimation should work. A point estimate should not be used for an uncertain parameter, nor should an analysis be based on one value when a whole range of other values are only slightly less likely.

Limitations to Using Only the Discovery Sequence

The approach of using only the discovery sequence information in estimating undiscovered resources ignores much additional information about the play. How many dry holes have been drilled? What fraction of the play has already been explored? How well known is the geology of the play? How homogeneous is the play in a geological sense? Some of this information can be incorporated into the prior distributions on the parameters. For example, the prior on N should reflect the fraction of the play area that has already been explored and the number of dry holes. However, it is more desirable to have an approach that takes all of this information into consideration directly in making estimates of undiscovered pool sizes. Such an approach is a logical next step from the approach given in this paper.

ACKNOWLEDGMENT

This work was supported by National Science Foundation contract ISI-8501001.

REFERENCES

Grace, J. D., 1986, Advantages and limitations of discovery process modeling—The case of the Northern and West Siberian Gas Plays: presentation at NATO Advanced Study Institute on the Statistical Treatment for Estimation of Energy and Mineral Resources, Lucca, Italy.

Kaufman, G. M., Balcer, Y., and Kruyt, D., 1975, A probabilistic model of oil and gas discovery, *in* J.D. Haun (Ed.), Estimating the volume of undiscovered oil and gas reserves, Studies in geology, vol. I: Oklahoma, American Association of Petroleum Geologists, p. 113–142.

Lee, P. J., and Wang, P. C. C., 1983a, Probabilistic formulation of a method for the evaluation of petroleum resources: Math. Geol., v. 15, p. 163–181.

Lee, P. J., and Wang, P. C. C., 1983b, Conditional analysis for petroleum resource evaluation: Math Geol., v. 15, p. 353–365.

Lee, P. J., and Wang, P. C. C., 1985, Prediction of oil or gas pool sizes when discovery record is available: Math. Geol., v. 17, p. 95–113.

Lindley, D. V., Tversky, A., and Brown, R. V., 1979, On the reconciliation of probability assessments: J. R. Statist. Soc. A. 142, Part 2, p. 146–180.

Morris, P. A., 1977, Combining expert judgments—A bayesian approach: Mgmt. Sci., v. 23, p. 679–793.

Savage, L. J., 1971, Elicitations of personal probabilities and expectations: J. Am. Statist. Assoc., v. 66, p. 783–801.

Spetzler, C. S., and Stael Von Holstein, C. S., 1975, Probability encoding in decision analysis: Mgmt. Sci., v. 22, p. 340–358.

Stone, L. D., 1983, The process of search planning: Current approaches and continuing problems: Operations Research, v. 31, p. 207–293.

Winkler, R. L., 1981, Combining probability distributions from dependent information sources: Mgmt. Sci. v. 27, p. 479–488.