Sankhyā: The Indian Journal of Statistics
1997, Volume 59, Series B, Pt. 3, pp. 288-312

STOCHASTIC MODELLING AND FORECASTING OF DISCOVERY, RESERVE AND PRODUCTION OF HYDROCARBON - WITH AN APPLICATION

By J.K. GHOSH, P. MAITI, A.C. MUKHOPADHYAY
and
M. PAL
Indian Statistical Institute, Calcutta

SUMMARY. The primary objective of this paper is to present some results in connection with forecasting of discovery and production of hydrocarbon in a partially explored basin. The focus in this article will be on stochastic modelling for the purpose of discovery, based on the idea of subjective probability or super population. The results of a particular basin in India have been presented here.

1. **Introduction**

At the request of the Oil and Natural Gas Commission (ONGC) of India, the Indian Statistical Institute evaluated the economic and physical consequences of various strategies for action in different oil basins in India. Roughly speaking, a strategy was an allocation of fund to different basins and, within a basin, to different activities like exploration, development, production and so on. The physical consequences were assessment of reserve and actual course of production. One of the basic questions asked was - 'how much oil would be discovered, if on the basis of existing geophysical evidence about a basin, so many wells were drilled or so much fund was spent on drilling at suitable sites, technically called prospects? A similar question relating production to the availability of funds over a time frame was asked. A preliminary account of the work is available in the report entitled ISI-ONGC collaborative research project (1990). The focus in the present article will be on stochastic modelling for the process of discovery, based on the idea of subjective probability or super population. The discovery model considered is not new, but many of the details in the application are. These include the methods of simulation and stochastic formulation of the basic

Paper received. November 1994; revised December 1997.
AMS (1991) *subject classification.* 62P99.
Key words and phrases. Stochastic, modelling, forecasting, empirical Bayes method, hydrocarbon.

questions asked. We will also present the results obtained for a particular basin, which will remain unidentified for reasons of confidentiality, and compare them with the results obtained in the report mentioned by non-stochastic regression-based econometric methods. This article along with the ISI-ONGC Collaborative Research Project Report (1990) will provide a complete picture of the work at the Indian Statistical Institute in this connection.

In Section 2, we present the model and pose the original question of the ONGC in the framework of the model. A very brief and simplified account of the physical aspects relevant to the formulation of the model is only provided. We also briefly review many other models that have been proposed in the literature, and discuss how some of our questions could be answered on their basis. In Section 3 is presented the methodology developed for providing answers to the questions with most of the technical details of proof omitted. In Section 4, we present results of a particular basin in India and compare with the results obtained earlier by the non-stochastic econometric method. In the final Section, a few unresolved technical questions relating to the model are raised.

2. Basic Model, Assumptions and Estimation

2.1 *Model with assumptions.*

We consider a partially explored basin for which geological and geophysical surveys are complete and exploration consists primarily in drilling wells in suitable sites decided upon from technical considerations. Exploratory wells will sometimes lead to the discovery of an oil bearing (economically feasible) field. Thus in a basin, we have a series of exploratory wells resulting in either success or failure. In case of a success, further suitable wells are drilled to delineate the field and a detailed technical appraisal is made to determine its size and other relevant parameters. The technical appraisal for an economically viable field also involves the preparation of an optimum production schedule to be followed in the field. The actual determination of the schedule is a lengthy and complicated process of physical simulation, which depends on besides size, many other geophysical characteristics of the field information which is collected on a regular basis from the exploratory and delineation wells. The schedule is reviewed and revised, if necessary, from time to time on the basis of the updated information collected from the development wells drilled for the purpose of production of oil from the field. Of course, it can be safely assumed that as soon as an exploratory well leads to a success, i.e., strikes oil, the process of delineating the field, leading to the determination of its size starts. The technical appraisal involving physical simulation, resulting in the projected optimum production cycle to be followed also follows immediately afterwards. But it may not be always most profitable or even convenient to start production as soon as these processes are complete in all such fields.

The discovery and production models along with associated costs are developed taking into account the discrete nature of discoveries, stochastic behaviour of the exploration activity and treating the lag between delineation of a field and the initiation of production from it as the only additional dicision variable of production along with the exploration schedule to be followed within a time frame.

The discovery model which follows has been considered by various authors since the pioneering work of Arps and Roberts (1958) and has been almost inalienably related to the problem of oil exploration in a basin and the rationale is well known. Special mention in this connection may be made of the following authors : Bloomfield *et al.* (1979), Drew *et al.* (1980), Kaufman *et al.* (1981), Smith *et al.* (1981), Gordon (1983a, 1983b), Nair and Wang (1989), West (1992). An asymptotic theory is provided by Bickel, Nair and Wang (1992).

A brief description of the assumptions of the model is provided below for the sake of completeness :

(i) The basin consists of a set of N fields which constitute a random sample from the super population field size distribution given by the p.d.f.

$$f(y \mid \theta), \quad y \geq 0.$$

The form of the p.d.f. is assumed *a priori*, with θ unknown. One may also think of $f(y \mid \theta)$ as a partly specified prior distribution with parameters to be estimated by some form of the empirical Bayes method. A fully Bayesian hierarchical method, which estimates errors in estimation and prediction better, would be our current preference. N may be treated as known, or unknown parameter or a random variable following a suitable probability distribution.

(ii) The chronological list of the results of exploratory wells drilled give rise to a 0-1 sequence, '0' denoting a failure and '1' denoting a success. The sequence exhibits statistical regularity in the sense that its behaviour can be explained in terms of a probability distribution.

(iii) Considering only the success points in (ii), we have a new sequence $\{y_n\}, n \geq 1$, where y_n represents the size of the nth discovery field. The sequence $\{y_n\}, 1 \leq n \leq N$ constitutes a realized sample, the drawing scheme being probability proportional to size without replacement (PPSWOR) from the population of N fields in the basin, the size measure for the unit value y_i being $w(y_i)$, a weight function depending on y_i.

Quite often $w(y_i)$ is taken as just y_i. The stochastic behaviour of the sequence $\{y_n\}, 1 \leq n \leq N$ is not dependent on the 0-1 sequence in (ii) in any other manner.

(iv) For the sake of simplicity, it is assumed that the optimum production cycle for a field can be predicted sufficiently accurately once its size is known. This cycle as estimated from the past data from the basin is also used for the fields to be discovered in future, assuming no change in technological and other environmental conditions having any bearing on the process.

Now, considering a finite population of N given units $(y_1, y_2, \ldots, y_N)$, the joint probability of the first n items drawn being $(y_1, y_2, \ldots, y_n)$ in order under PPSWOR is

$$P(y_1, y_2, \ldots, y_n \mid y_1, y_2, \ldots, y_N) = \prod_{i=1}^{n} \frac{w(y_i)}{t + b_i} \qquad \ldots (2.1)$$

where $t = \sum_{i=1}^{N} w(y_i) - \sum_{i=1}^{n} w(y_i), \quad b_i = \sum_{j=i}^{n} w(y_j).$

Under the super population model with N fixed, as spelt out in the assumption, the unconditional distribution of $(y_1, y_2, \ldots, y_n)$ can be obtained by summing over all possible values $\{y_1, y_2, \ldots, y_N\}$, multiplying by the joint density function and integrating over the unobserved values i.e., the probability density function of $(y_1, y_2, \ldots, y_n)$ is

$$\begin{aligned}
& P(y_1, y_2, \ldots, y_n; \theta, N) \\
&= k\pi_{i=1}^{n} w(y_i) f(y_i \mid \theta) \\
&\quad \times \int \ldots \int \sum_{i=1}^{n} \frac{\frac{1}{b_i + t}}{\pi_{(j \neq i)=1}^{n} (b_i - b_j)} f(y_{n+1}) \ldots f(y_N) \pi_{i=n+1}^{N} dy_i \\
&= k\pi_{i=1}^{n} w(y_i) f(y_i \mid \theta) \\
&\quad \times \int \ldots \int \sum_{i=1}^{n} A_i \left(\int_0^\infty e^{-\lambda(b_i + t)} d\lambda \right) f(y_{n+1}) \ldots f(y_N) \pi_{i=n+1}^{N} dy_i \\
&= k\pi_{i=1}^{n} w(y_i) f(y_i \mid \theta) \sum_{i=1}^{n} A_i \qquad \ldots (2.2) \\
&\quad \times \int_0^\infty e^{-\lambda b_i} \left(\int \ldots \int e^{-\lambda w(y_{n+1} - \ldots - \lambda w(y_N)} \pi_{i=n+1}^{N} f(y_i) dy_i \right) d\lambda \\
&= k\pi_{i=1}^{n} w(y_i) f(y_i \mid \theta) \sum_{i=1}^{n} A_i \int_0^\infty -e^{-\lambda b_i [L(\lambda|\theta)]^{N-n}} d\lambda \\
&= k\pi_{i=1}^{n} w(y_i) f(y_i \mid \theta) \sum_{i=1}^{n} \int_0^\infty A_i \; e^{-\lambda b_i} [L(\lambda \mid \theta)]^{N-n} d\lambda
\end{aligned}$$

where $k = \dfrac{N!}{(N-n)!}$, $A_i = [\pi_{j(\neq i)=1}^{n} (b_i - b_j)]^{-1}$ and $L(\lambda \mid \theta) = \int_0^\infty e^{-\lambda w(y)} f(y(\theta)\; dy$. The expressions (2.1) and (2.2) can be obtained easily from the comparable expressions given in Kaufman *et al.* (1975) and West (1992).

From, (2.2), the conditional distribution of y_{n+1} given $y_1, y_2, \ldots, y_n$ is given by the p.d.f.

$$\begin{aligned} & P(y_{n+1} \mid y_1, y_2, \ldots, y_n; \theta, N) \\ = \quad & (N-n) w(y_{n+1}) f(y_{n+1} \mid \theta) \\ & \times \textstyle\int_0^\infty \{ \sum_{i=1}^{n} (-\frac{A_i}{b_i}) e^{-(b_i + w(y_{n+1}))\lambda} \\ & + \frac{1}{b_1 b_2 \ldots b_n} \} e^{-\lambda w(y_{n+1})} . [L(\lambda \mid \theta)]^{N-n-1} d\lambda \\ & \times \ [\sum_{i=1}^{n} A_i e^{-b_i \lambda} [L(\lambda \mid \theta)]^{N-n} d\lambda]^{-1} \end{aligned} \quad \ldots (2.3)$$

The expression (2.3) can be found in Nair and Wang (1989).

In both the expressions (2.2) and (2.3), N can be assumed to follow a probability distribution, say a truncated poisson distribution and the corresponding densities can be easily obtained by incorporating this distribution in the expressions.

In the model and subsequent methods which have been applied to the data from a partially explored basin, N is treated as fixed and known, i.e., in addition to assuming N to be non-random, we do not plan to estimate it from the likelihood (2.2) given the data $(y_1, y_2, \ldots y_n)$. Instead, N has been estimated from the estimate of the success ratio and the number of prospects determined through technical appraisal reports. Estimation of N has been discussed in section 2.3.3. It may be noted that the expert views were very much consistent with this estimate of N.

The widely accepted assumption in describing the discovery process model is that of 'larger sizes have the higher probability of being discovered first'. Some authors also feel that discovery probabilities might be related to some monotonically increasing function of size. (Carrol and Smith (1980); Bloomfield *et al.* (1979); Smith and Ward (1981); Lee and Wang (1983b); Nair and Wang (1989); West (1992)). Therefore a general framework for size biased sampling supposes that units are selected with probability proportional to some non-negative function $w(.)$, increasing in the size of the field.

A typical choice for $w(.)$ by all the authors is $w(y) = y^\alpha$ for some specified constant $\alpha \geq 0$. The power of proportionality α is also known as the power of the discoverability. $\alpha = 0$ represents the situation in which fields have equal discoverable probabilities regardless of size and $\alpha = 1$ implies the discovery probabilities are strictly proportional to the size of the field.

Interpretation for different values of α. Let R_E be the cumulative discovery corresponding to the cumulative cost of exploration. The graphs of R_E *vs.* C_E relationship should look as follows for different possible values of α.

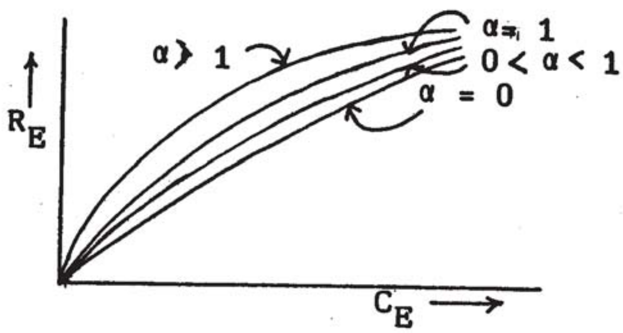

Figure 1 : Cumulative expected reserve as a function of cumulative cost

Observe that in the Figure 1 we have actually shown the cumulative expected reserve against cumulative cost. But the data have jumps because discoveries are made at discrete points. In that case the R_E *vs.* C_E graph should look like a step function as shown below :

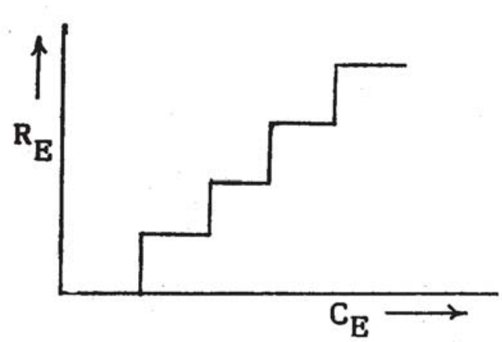

Figure 2 : Cumulative actual reserve against cumulative cost

It is felt that $w(y)$ proportional to y itself will mean that relationship on cumulative reserve and cumulative cost of exploration will appear as follows :

$$R_E = k - a(b - C_E)^r, \qquad 0 < C_E < b,$$

at least for C_E close to b. Thus if this relationship cannot be fitted to the data in a precise manner then this probably would indicate that $w(y)$ is not proportional to y itself, but to some power y^α, $0 < \alpha < 1$ or some other function of y.

With initial value of $k = 708.58$ million metric ton (mmt) and $b = 1936.83$ Rs. million, the fitted relationship was

$$R_E = 708 - 0.05781(1935.831 - C_E)^{1.2057}$$

with correlation between observed R and fitted R (i.e., $\rho(R, \hat{R})$) equal to 0.942. As we increase b up to 6436.83 by 100 units keeping k at 708.58 million metric ton, $\rho(R, \hat{R})$ keeps on increasing to 0.9595. Also a approaches to zero and r to 7.9028.

We have also plotted the step functions of cumulative of a power (α) of reserve against cumulative cost for different values of α taking $\alpha = 0.2$, 0.5, 0.8, 1.0, 1.2, 1.5 & 1.8. All the diagrams look similar. The graphs do not show any perceptible change in the shape. Thus from these analyses we can not conclude anything about the value of α especially whether $\alpha = 1$ or not.

We have taken $w(y) = y$ for all further developments. Empirical evidence on the deposit size distributions indicate that unimodal and positively skewed with a flat-right-tail are the characteristic traits of the relative frequency histogram of size - areas, rock volume, hydro carbon in place or recoverable hydrocarbon deposits. About the form of $f(y \mid \theta)$, Lognormal and Gamma have been found to be two competitors.

Actual numerical computations of the expressions were tried on VAX 781 at Keshab Dev Malaviya Institute of Petroleum Exploration (KDMIPE), Dehradun using the observed field size data under different plausible assumptions regarding $f(y \mid \theta)$. Expressions for the likelihood for different values of the sets of the parameters could be calculated sufficiently accurately (with double precision) for all the reasonable and practically realizable assumptions about the form of field size distributions, e.g., Gamma, Lognormal for all n and $N \leq 10$.

But for larger values of N and the available field size data, the computational results were found to behave erratically and hence their reliability was questionable. Since the data obtained systematically indicated that the number of fields discovered upto 1985 was far greater than 10, computational intractability posed the greater hurdle in obtaining the maximum likelihood estimates of the parameter. Infact, all the applicable and approximate methods of numerical approximation and quadrature formula were tried, but not much could be achieved.

The form of $f(y \mid \theta)$ is taken to be Gamma with $\theta = (\alpha, \beta)$ for all subsequent computations which follow. The point which can be stated in justification is that the hypothesis of a lognormal distribution which happens to be widely accepted in this context is known to be hardly distinguishable from an appropriate Gamma distribution, unless the sample size is quite large.

In the framework of the model discussed and under assumptions stated, the questions asked by the ONGC are as follows :

1. Given $y_1, y_2, \ldots, y_n$, the $0-1$ sequence up to the nth discovery and relevant cost parameters, what is the additional expected cost (with standard deviation) of realizing an expected additional discovery, $E(y_{n+1} \mid y_1, y_2, \ldots, y_n)$ and from this what is the additional expected cost (with standard deviation) of establishing a unit of additional expected discovery in the basin?

2. Given $y_1, \ldots, y_n$, the $0-1$ sequence up to nth discovery and relevant production and cost parameters, what is the additional expected cost (with standard deviation) of making an additional unit of expected production of oil from the basin.

3. Given $y_1, \ldots, y_n$, the $0-1$ sequence up to the nth discovery, the lag parameters and optimum production solely depending on field sizes as pointed out in assumption (iv) in Section 2.1 and the exploration schedule to be followed within a time frame, what will be expected cumulative cost curve $C(t)$ and the expected cumulative production curve $P(t)$ for the basin over time t (measured from the time of the nth discovery)?

Changing the decision variables within the overall budgetary and technical restrictions which appear realistic, different scenarios of the production and cost profiles given by the $P(t)$ and $C(t)$ curves can be created and presented to the appropriate people for final decision.

The principal objective of the project was the development of methodologies needed for the purpose.

2.2 *Review of available techniques.*

2.2.1 *Stochastic modellings.* On the basis of the same set of assumptions as in the present article, Nair and Wang (1989) gave a different presentation of (2.1), i.e., the likelihood for θ given the data. The log of the likelihood is shown to be

$$\text{Log } L(\theta) = \text{constant} + \sum_{j=1}^{n} \log f(y_j \mid \theta) + \log S(\theta) \qquad \ldots(2.4)$$

where, $S(\theta) = \int_0^\infty [L(t \mid \theta)]^{N-n} g_n(t)dt$, $g_n(t)$ = general gamma density given by

$$\sum_{i=1}^{n} A_i[b_i e^{-tb_i}], \quad t > 0$$

and $L(t \mid \theta), b_i$'s and A_i's are as defined in (2.2).

From (2.4), they found analogous to (2.3), the conditional distribution of $y_{n+1}, y_{n+2}, \ldots, y_N$, given the data as

$$P(y_{n+1}, y_{n+2}, \ldots, y_N \mid y_1, y_2, \ldots, y_n, \theta, N) = \int_0^\infty \pi_{i=1}^n K(y_i \mid t, \theta) h_\theta(t) dt \qquad \ldots(2.5)$$

where, $h_\theta(t) = [L(t \mid \theta)]^{N-n} g_n(t)/S(\theta)$ and

$$K(y \mid t, \theta) = exp\ (-tw(y)) f(y \mid \theta)[L(t, \theta)]^{-1}.$$

The distribution (2.5) is used to obtain the expectation of $Z = y_{n+1}$ given $y_1, y_2, \ldots, y_n$. Finally, they showed that the expression used for the maximisation step of EM algorithm is

$$Q(\theta \mid \theta^*) = \sum_{i=1}^{n} log\ f(y_i \mid \theta) + (N-n) E_{\theta^*}(log\ F_\theta(z) \mid data).$$

The underlying principle of work due to Mike West (1992) stems from simulating the posterior distribution of θ given the data. His is a fully Bayesian treatment and has the advantage that his method of simulation (Gibbs Sampling) can handle the problem depending on many variables including size. However, our method of simulation to be discussed in the following section may take much less time and so may be better in the univariate case.

In respect of the size distribution, Bickel *et al.* (1992) assumed that $z_1, z_2, \ldots, z_k$ are the distinct values of field size with multiplicities $N_1, N_2, \ldots, N_k$ respectively. The unknown parameters are $(N_1, N_2, \ldots, N_k)$ and the log likelihood of $(N_1, N_2, \ldots, N_k)$ is given by $L(N)$, where

$$\begin{aligned} e^{L(N)} &= \prod \frac{N_k!}{(N_k - n_k)!} \prod_{i=1}^{n} \frac{w(y_i)}{\sum N_r w(y_r) - D(i)} \\ D(i) &= \sum_{j=0}^{i-1} w(y_j)\ \ with\ \ w(y_0) = 0,\ \ w(y_j), \quad j = 1, 2, \ldots, n \end{aligned} \qquad \ldots (2.6)$$

as defined in (2.1). The sampling scheme assumed is the same as in (2.1).

The problem of estimation of $R = \sum_{k=1}^{K} z_k N_k$ in their model is same as the problem of estimating the frequency distribution $\{N_k\}, k = 1, 2, \ldots K$ and $N = \sum_{i=1}^{K} N_i$. Since obtaining MLE's for $(N_1, N_2, \ldots, N_k)$ appeared quite different under the model (2.6), they considered the following asymptotic structure.

(a) As $N \uparrow \infty, K =$ the number of distinct values is fixed and the distinct values $(z_1, z_2, \ldots, z_k)$ converge to some fixed finite values.

(b) $\nu_k (\equiv N_k/n)$ satisfies $0 < \overset{limit}{N \uparrow \infty}\ \ \nu_k$ for $k = 1, 2, \ldots, K$ and $f(\equiv n/N)$ satisfies $0 < \overset{limit}{N \uparrow \infty}\ \ f < 1$.

2.2.2 *Econometric approach.*

Attempts in explaining the discovery process following econometric methods have been made, among many others, by Newendrop (1976), Mc Cray (1975), Ramsey (1981), Khazoom (1971). Errickson and Spann (1971), Epple and Hansen (1979) who considered cost implications for modelling exhaustible resources supply in an econometric framework. In the I.S.I. - ONGC collaborative research project report (1990), an in depth study was made following the econometric approach along with the approach on stochastic modelling. Incidentally, the econometric approach was a regression approach where the model was developed basically in terms of relationships between accumulated volume of discovery like exploratory drillings, delineation drillings and geological and other types of surveys. For this purpose, three different kinds of data arrangement were made, namely discoverywise data, yearwise data and equal cost interval data. Different alternative algebraic forms were estimated on the basis of actual data in order to identify the best fitting functional forms for the reserve/delineation cost function for each type of data set. Finally, a modified exponential form of reserve function $\hat{R} = K - ab\, C^E$ was found to behave well for each type of data, although, the estimates of parameters differed to some extent for the different types of data sets. For the application problem to be considered in Section 4, the estimates of K, a and b found are as given in the following table.

Table 2.1. ESTIMATION OF K, a AND b

sr. no.	type of data	estimates of parameters			
		K	a	b	$r_{R\hat{R}}$
1	Discoverywise	741.55	590.25	0.9066	0.968
2	Equal Cost interval	1167.16	950.34	0.9704	0.977
3	Yearwise	711.24	651.17	0.8728	0.960

In the table above, r denotes the correlation between the actual values and the estimates. The data set on discoveries and size are provided in the Table 5.1. An idea about the cost parameter can be had from the discussions in Section 4.

Following the same rationale as in the case of discovery model, a production model can be developed by estimating the functional relationship between the level of production Y and the relevant dependent variables. For details, one is referred to the report, viz., I.S.I. - ONGC collaborative research project (1990).

2.3 *Estimation procedure*

2.3.1 *Parameters of the size distribution.* Given N and $y_1, y_2, \ldots, y_n$ in order, the first problem is to estimate the parameters α and β of the Gamma size distribution assumed. The procedure is basically the method of moments. On algebraic simplification, one can show the following :

$$E(y_i) = \frac{\beta^i(\beta+1)}{\alpha} \times \frac{N(N-1)\ldots(N-i+1)}{(N\,\beta+1)(\overline{N-1}\,\beta+1).(\overline{N-i+1}\,\beta+1)}$$
$$i = 1, 2, \ldots, n$$

$$
\begin{aligned}
E(y_i y_j) &= \frac{\beta^j(\beta+1)^2}{\alpha^2} N(N-1)\ldots(N-j+1) \times \{(N\,\beta+2)(\overline{N-1}\,\beta+2) \\
&\quad \ldots(\overline{N-i+1}\,\beta+2) \times (\overline{N-1}\,\beta+1)(\overline{N-2}\,\beta+1) \\
&\quad \ldots\overline{N-i+1}\,\beta+1\}^{-1}, \\
E(y_i^2) &= \frac{\beta^i(\beta+1)(\beta+2)}{\alpha^2} N(N-1)\ldots(N-i+1) \\
&\quad \times\{(N\,\beta+2)(\overline{N-1}\,\beta+2)\ldots(\overline{N-i+1}\,\beta+2\}^{-2}, \quad i=1,2,\ldots n
\end{aligned}
$$

One can now equate

(a) $\sum_{i=1}^{n} y_i$ to its expected values $E(\sum_{i=1}^{n} y_i)$

(b) $\sum_{i=1}^{n} y_i^2$ to its expected value $E(\sum_{i=1}^{n} y_i^2)$

in order to estimate α and β.

Actually in a practical problem, initial estimates of α and β can be obtained by assuming the sample to be a simple random sample and then Newton-Raphson or any other appropriate method of iteration can be used to solve the equations in (2.7).

After the estimation of parameters, it is needed to verify the justifiability of PPSWOR assumption particularly as stated in 2.1. This is done as follows :

Substituting the estimates of α, β for α, β respectively one can get appropriate estimates of $E(y_i)$ and $var(y_i) = E(y_i^2) - (E(y_i))^2$ for all $i = 1, 2, \ldots, n$. Then one can use a control chart type technique with the

Central line as $E(y_i)$

Lower control limit as $E(y_i) - 2\sqrt{var(y_i)}$

and

Upper control limit as $E(y_i) + 2\sqrt{var(y_i)}, i = 1, 2, \ldots, n$

when observed y_i's are plotted in the chart, we expect all or almost all the observations within the control limits.

In this connection it may be pointed out that one could also use a more elaborate method of maximum likelihood estimation for parameters via EM-algorithm based on (2.2) as developed by Nair and Wang (1989), where they use the conditional distribution (2.3) for solving the expected step of the algorithm. From a Bayesian point of view, this would be I.J. Good's ML II approach, vide Berger (1985, p.99). The method of moments used by us, even if it is less efficient, is fairly straightforward. Obtaining large sample variance of the estimates by this method would also be fairly easy, in case one is interested. On the other hand, the method of obtaining ML estimates given by Nair and Wang (1989) is definitely computationally more involved.

2.3.2 *Parameters of the probability distribution governing the 0-1 sequence.* Regarding the 0-1 sequence as stated in the assumption (ii) of section 2.1, let W_n represent the number of wells drilled following the nth success and up to and including the next.

It is the probability distribution of W_n for all $n \geq 1$ which is crucial to the computation of exploration and delineation cost for discovery.

From the partial data on the 0-1 sequence given, the appropriate distribution of W_n for different values of n are arrived at and used for subsequent projections for the applications considered in section 4. The appropriate distribution of W_n is taken to be geometric on the evidence of the available data subjected to relevant statistical tests.

Based on the 0-1 sequence and the number of prospects determined through the results of the technical appraisal reports, i.e., on the basis of the success ratio and of the number of prospects determined, N was estimated as $N =$ (No. of prospects/$\hat{P}$) = 40. This estimate was quite consistent with the expert views.

Theoretically it is possible to endogenise the variable N and obtain expressions for its analytical solution under our set up. But the computational facility at our disposal proved insufficient for getting numerical solutions to the analytical expression. The same was experienced by Nair and Wang (1989) and their analysis was also based on a known population size N. The difficulties are very much inherent in the problem of simultaneously estimating θ and N. On the other hand, in the oil reserve context, there is a degree of externally available expert opinion and data based geological/geophysical information could be used. We therefore treated N as an exogeneous variable and solved the model for a specified value of N. However, in section 4, sensitivity of project reserve, production with respect to the total number of fields has been carried out.

3. Simulation Techniques

3.1 *Discovery model.*

3.1.1 *Cost estimation.* Given W_n and y_{n+1}, the additional cost needed to make an additional discovery of size y_{n+1} in the $(n+1)th$ discovery is given by

$$T_{n+1} = c_1 W_n + c_2(y_{n+1})$$

where c_1 = cost of an exploratory well and $c_2(x)$ = a linear function of x gives the cost of delineation of a field of size x, in a basin, based on technical appraisal results.

If we know how to calculate $E(W_n)$ and $E(y_{n+1} \mid y_1, \ldots, y_n)$ = the expected additional discovery, the conditional expectation of $c_2(y_{n+1})$ given $y_1, y_2, \ldots, y_n$ can be calculated at least approximately. The conditional variance can be similarly computed once we know how to calculate $var(W_n)$ and $var(y_{n+1} \mid y_n)$.

3.1.2 *Conditional distribution of* y_{n+1} *and simulation procedure.*

3.1.2.1 Simulation 1. We make the simplifying assumption

$$\begin{aligned} E(y_{n+1} \mid y_1, y_2, \ldots y_n) &= \delta_0 + \delta_1\, y_1 + \cdots + \delta_n\, y_n \\ &= a \text{ linear function of } y_1, \ldots, y_n \end{aligned}$$

and

$$V(y_{n+1} \mid y_1, y_2, \ldots y_n) = \sigma^2 \text{ (constant).} \qquad \ldots(3.1)$$

Under the assumption (3.1), the parameters can be estimated from the equations

$$\begin{aligned} E(y_i\, y_{n+1}) &= \delta_0 E(y_i) + \delta_1 E(y_1\, y_i) + \ldots + \delta_n E(y_n\, y_i) \\ & \qquad i = 0, 1, 2, \ldots, n; \quad y_0 = 1 \text{ identically} \end{aligned} \qquad \ldots(3.2)$$

and

$$\sigma^2 = V(y_{n+1})(1 - \rho^2_{n+1.1,2,\ldots,n}) = \frac{\mid D(y_1, y_2, \ldots y_{n+1}) \mid}{\mid D(y_1, y_2, \ldots y_n) \mid} \qquad \ldots(3.3)$$

where $\rho_{n+1.1,2,\ldots,n}$ is the multiple correlation coefficient of y_{n+1} on $y_1, y_2, \ldots y_n$ and $D(y_1, y_2, \ldots, y_n)$ indicates the dispersion matrix of the random variables $(y_1, y_2, \ldots, y_n)$.

We assume the conditional distribution of y_{n+1} given $y_1, y_2, \ldots, y_n$ to be Gamma with mean given by (3.1) and variance given by (3.2).

It may be pointed out that determining $E(y_{n+1} \mid y_1, y_2, \ldots, y_n)$ and $var(y_{n+1} \mid y_1, y_2, \ldots, y_n)$ in the manner described above is quite simple. What is more important is to examine how far the approximate distribution thus obtained gives a good fit to the exact distribution in (2.3). For this purpose, extensive numerical exercises were carried out on the basis of available field size data for the basin. From amongst the discovered fields, a sample of the first $N(\leq 10)$ fields was obtained to constitute a hypothetical basin for which the conditional distributions were obtained for different values of n both by the exact formula and by the approximate method mentioned above. The same exercise was repeated for quite a few such samples of different sizes, i.e., values $N(\leq 10)$. The numerical computations showed that the estimates of the conditional p.d.f. obtained by the methods are remarkably close. These encouraging results prompted us to accept the approximate method as a practical procedure for determining the conditional p.d.f. of y_{n+1}, given $y_1, y_2, \ldots, y_n$.

We give the steps involved in the simulation based on this approximation:

(i) obtain an observation from the approximate Gamma distribution $P(y_{n+1} \mid y_1, \ldots, y_n; \theta, N)$ and call it y_{n+1};

(ii) obtain an observation from W_n;

(iii) calculate $c_2(y_{n+1})$;

(iv) Treat y_{n+1} as the realised $(n+1)th$ field size and repeat the steps (i) through (iii) with n replaced by $n+1$.

The whole sequence of operations can be repeated as many times as one likes and the relevant expectations with associated standard deviations can be computed.

Simulation II. Let $z_1, z_2, \ldots z_{N-n}$ be the undiscovered magnitudes and $y_1, y_2, \ldots y_n$ as usual the discovered magnitudes with respect to the sequence of discovery.

Let $R = z_1 + z_2 + \ldots + z_{N-n} =$ the reserve yet to be discovered in the basin. The following facts are noted.

(a) Given $y_1, \ldots y_n$ and R the conditional density of $z_i/R = U_i$, $i = 1, 2, \ldots,$
$N-n$ is Dirichlet with parameters all equal to β, i.e., U_i's are independent of

$$y_1, \ldots, y_n \text{ and } R. \qquad \ldots(3.4)$$

(b) Conditional density of R given $y_1, \ldots y_n$ is Const. $e^{-\alpha R} R^{(N-n)\beta-1}.h(R)$, where

$$h(R) = \prod_{i=1}^{n} \pi_i, \pi_i = \frac{y_i}{y_i + y_{i+1} + \ldots y_n + R}, i = 1, 2, \ldots, n \qquad \ldots(3.5)$$

To estimate $E(R_k \mid y_1, \ldots, y_n)$, where R_k is the sum of the next k discoveries, $1 \leq k \leq N-n$, $R_{N-n} = R$, the following simulation procedure is suggested :

(i) Generate M sets of $(N-n)$ i.i.d. Gamma (α, β) variates as $Z_1^{(i)}, Z_2^{(i)}, \ldots, Z_{N-n}^{(i)}$, $i = 1, 2, \ldots, M$.

(ii) $R^{*(i)} = \sum_{j=1}^{N-n} Z_j^{(i)}$ is a sample from $(G(\alpha, \overline{N-n}\beta))$ and $Z_j^{(i)}/R^{*(i)} = U_j$, $j = 1, 2, \ldots, N-n-1$ is a sample from Dirichlet distribution described in (a) above, $i = 1, 2, \ldots, M$.

(iii) Given $Z_1^{(i)}, Z_2^{(i)}, \ldots, Z_{N-n}^{(i)}$, draw the observations one by one, using PPSWOR sampling scheme with size measure as the value of the variable. Let the sequence of observations be denoted by $w_1^{(i)}, w_2^{(i)}, \ldots, w_{N-n}^{(i)}$.

(iv) Let $V_k^{(i)} = w_1^{(i)} + w_2^{(i)} + \ldots + w_k^{(i)}$, $1 \leq k \leq N-n$; $i = 1, 2, \ldots, M$.

(v) An estimate of $E(R_k \mid y_1, \ldots, y_n)$ is provided by

$$\frac{\frac{1}{M}\sum_{i=1}^{M} V_k^{(i)} h(R^{*(i)})}{\frac{1}{M}\sum_{i=1}^{M} h(R^{*(i)})}. \qquad \ldots(3.6)$$

One should vary M to see where stability is attained.

The estimator (3.6) converges to $E(R_k \mid y_1, \ldots, y_n)$. For estimating $var(R_k \mid y_1, \ldots y_n)$ a similar procedure can be adopted.

In both the procedures, simulation I and simulation II, we are trying to approximate an integral by some intuitive technique which makes the required computation very simple. It may be noted in this connection that Nair and Wang (1989), while developing an EM algorithm for finding the ML estimators of the parameters of population size distribution, given the data, have to actually evaluate this conditional distribution numerically, which could be adopted as it is for our purpose of simulation. However, the computation as pointed out in connection with the estimation procedures in section 2.3.1 would have been much more involved.

Unlike the PPSWOR scheme in survey sampling, we assume that size measure is a function of the unknown population values. The model (2.1) has been found to be useful in incorporating the size-bias inherent in many types of discovery data. Our estimating procedures described above would be applicable to many other problems with size bias sampling.

3.2 Production model.

For this model, time is taken to be discrete for the sake of simplicity. The present moment is taken as '0'. The fields in a basin at any point of time can be classified into the following 4 categories:

(A_1) : Set of fields where production is taking place currently;

(A_2) : Set of fields for which delineation has been completed, but production is yet to start;

(A_3) : Set of fields discovered but for which delineation is not yet completed;

(A_4) : Set of fields to be discovered in future.

For field j in the basin, let d_j represent the lag between the completion of delineation and beginning of production. Here d_j is a decision variable.

For a field j, there is an optimum production cycle depending on the size, like

$$P_j(t), P_j(t+1), \dots, P_j(t+l_j)$$

with associated costs as

$$C_j(t), C_j(t+1), \dots, C_j(t+l_j),$$

assuming that production starts at a point of time t and l_j = maximum length of life of a field j after which production terminates.

Optimum production cycle is understood in the following way.

For a field whose size is estimated after the completion of delineation process, a production schedule is prepared based on technical and cost considerations. The production schedule is actually arrived at on the basis of an elaborate reservoir simulation study carried out for this purpose. While doing so, not only the size of the field, but quite a few other geological parameters like porosity, pressure level, viscosity etc., about which reliable information is collected during the completion of the discovery process for a field are considered and used in arriving at the so called optimum production and cost schedules for a field.

Let δ_j be the time of delineation required for field j and if the field belongs to A_3, let f_j represent the length of time for which delineation has been carried out already.

For a field j in category A_1, let x_j represent the period of time for which production has continued so far.

For a field j in category, A_4, we write t_j as the time point when the discovery occurs and y_j for its size (t_j's and y_j's are obtained by the technique of simulation I or II used for the discovery model).

The production density at time t from the basin, say $P(t)$, can be written as

$$P(t) = \sum_{j \in A_1} P_j(t+x_j) + \sum_{j \in A_2} P_j(t-d_j) + \sum_{j \in A_3} P_j(t-d_j-\delta_j+f_j) + \sum_{j \in A_4} P_j(t-t_j-d_j-\delta_j) \qquad \ldots (3.7)$$

Associated costs can be added to get the total production cost at time t. The total cost at time t, say $C(t)$ is the production cost + exploration cost. For details, the reader is referred to the report entitled "Estimation of Discovery and Production Costs of Hydrocarbon with Some Applications to Indian Data" (ISI-ONGC Collaborative Research Project (1990)).

The determination of t_j for field j in category A_4 is done in the following manner.

From the simulation model for discovery, we get for the fields chronologically in order : Future discovery size y_j, cumulative no. of exploratory wells needed for field j : n_j. Suppose an exploration strategy specifies N_k exploratory wells to be drilled at time k.

Let $N_0^* = N_0$ and $N_k^* = N_{k-1}^* + N_k, \ k = 1, 2, \ldots$ If $N_k^* \leq n_j < N_{k+1}^*$, we take t_j as the point of time of discovery of the j-th field, while δ_j's and d_j's are obtained from the available data on the known fields.

Now for a given exploratory strategy, a set of values of the decision variables and a set of t_j's and y_j's for fields in category A_4 to be obtained by the simulation techniques as explained, one can compute the projected production $p(t)$ at time t and the associated total cost $c(t)$. The simulation exercise can be repeated to get the expected quantities with associated standard deviation.

4. **Application**

The simulation techniques explained in section 3 were applied to the available data for different fields of India to get a basinwise and national level picture in this connection. Full details are to be found in the report already mentioned. Summary results of the application in one of the standard basins are reproduced hereunder.

For this basin under consideration data were available on 227 exploratory wells. The 0-1 sequence was extensively analysed and the hypothesis of Bernoulli model found to be acceptable on the basis of a large number of statistical tests which were applied to the data in different forms. The probability of success p was estimated to be 0.12.

Out of 227 exploratory wells, reserve was established for 63 fields (Table 5.1). Considering the sizes, 27 fields were found to be economically viable and hence in our analysis, we have taken 27 to be the number of discoveries.

The available data gave rise to the following figures and relations:

$$\begin{aligned} C_1(\text{Rs. million}) &= 8.1 \\ C_2(\text{Rs. million}) &= 64.1 + 3.1y \\ \hat{N} &= 40 \\ n &= 27. \end{aligned}$$

For the production simulation exercise, the following data were used in addition:

$$d_j = \begin{cases} 0 \; for \; y_j \geq 4 \text{ mmt} \\ 4 \; for \; y_j < 4 \text{ mmt}. \end{cases}$$

The $P_j(t)$ curve for field j in category 4 was also assumed to be of the type :

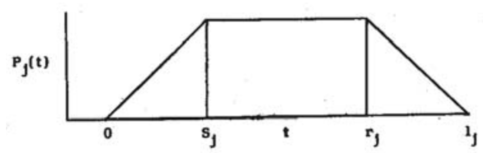

Figure 3 : $P_j(t)$ curve for a field in A_4

The values of S_j, r_j, l_j depending on size were based on the analysis of past data and expert opinions. The total area under the curve $P_j(t)$ is the recoverable reserve for field $j = y_j \times$ recovery factor. The recovery factor was appropriately estimated from the past data.

For a field already producing, the $P_j(t)$ curve was taken as:

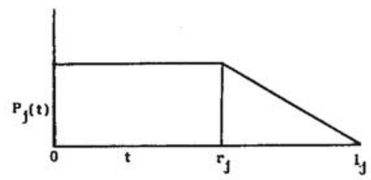

Figure 4 : $P_j(t)$ curve for a field in A_1

The exploration strategy utilised for the simulation result presented was :

	Time k										
	0	1	2	3	4	5	6	7	8	9	10
N_k	5	6	6	9	9	7	7	7	7	8	-

The summary results for the above strategy are presented in the following table along with comparable results for the econometric analysis. (The regression equations based on econometric methods are presented in Table (4.1).

Table 4.1 : ESTIMATES OF THE PARAMETERS OF POPULATION SIZE DISTRIBUTION

parameter	estimator
α	.0211132
β	.530576

Table 4.2 : ESTIMATE OF ULTIMATE EXPECTED RESERVE AND EXPECTED DISCOVERED RESERVE ON FIRST FIVE DISCOVERIES

method of simulation	ultimate reserve (mmt)	reserve for first five discoveries
Approximation I	757	55.7
Approximation II	826	80.0
Econometric Methods (Yearwise)	711	−−
Econometric Methods (Discoverywise)	742	−−

The above figures for ultimate reserve may be compared with two naive estimates. The first naive estimates is

$$NE_{\hat{\alpha}\,\hat{\beta}}(z) = 40 \times 25.10638 = 1004.26\, mmt$$

The Second naive estimate is

$$\begin{aligned} & y_1 + y_2 + \cdots + y_n + (N-n)E_{\alpha\,\beta}(z)\ mmt \\ = \ & [678.33 + (40-27)E_{\alpha\,\beta}(z)]\ mmt \\ = \ & 1004.7129\, mmt. \end{aligned}$$

Comparing the ultimate reserves it is found that $\hat{N}$ happens to be an over estimate. Both these figures tend to overestimate because they ignore the fact that probability of selection is proportional to size. Only relatively small fields may remain undiscovered.

The following two tables represent production and cost profiles corresponding to the following exploration strategy.

	Time k									
	0	1	2	3	4	5	6	7	8	9
N_k	5	6	6	9	9	7	7	7	7	8

The results in the Table 4.3 are due to the method of simulation (Approximation I), whereas for the sake of comparison, the projected production and cost figures obtained by econometric method (under the same above strategy) are displayed in the Table 4.4.

Table 4.3 : PROJECTED PRODUCTION AND COST PROFILES FOR THE TIME POINTS 0 TO 9 (METHOD OF SIMULATION, APPROXIMATION I), '0' REPRESENTING FOR THE YEAR 1985 FOLLOWED BY 1, 2, ..., 9 FOR THE YEARS 1986, 1987,... AND SO ON.

t	E (P(t))	E (C(t))	S.D. (C (t))
	mmt	Rs. million	
0	6.28	772.0	0
1	6.24	786.2	21.7
2	6.22	802.4	27.9
3	6.53	884.9	31.9
4	6.72	913.4	40.6
5	6.87	926.8	45.2
6	6.99	922.3	52.1
7	7.12	950.9	57.6
8	7.29	980.5	57.0
9	7.50	985.5	67.5

Table 4.4 : PROJECTED COST AND PRODUCTION FIGURES BASED ON THE ECONOMETRIC ANALYSIS

period	production (mmt)	cost (Rs. million)
0 - 4	8.7366	2871
5 - 9	7.7957	2614

mmt : million metric ton

4.1 *Sensitivity of project reserve, production with respect to the total number of fields.*

The models presented involve a number of variables one of which is the number of fields in a basin (N). While describing the model, it was mentioned that though theoretically, it is possible to endogenise the variable N and obtain expression for its analytical solution under our set up, the computational facility at our disposal proved insufficient for getting numerical solutions to the analytical expressions. We were therefore forced to treat N as an exogenous variable and solved the model for specified values of N. With the help of the fastest computers available in the world now, it may be possible to get the numerical solutions.

For this purpose, the simulation exercise was carried out for two different values of N, namely 40 and 60 and some related results are presented in the following tables.

Table 4.5 : Projected production, production cost

$N = 40$

item	1985-86	1986-87	1987-88	1988-89	1989-90	1990-91
1. No. of exploratory wells	16	17	18	20	22	25
2. Exp. discovery (mmt)	0.0	23.72	22.02	14.97	9.42	5.65
3. Exp. total production (mmt)	6.28	6.24	6.22	6.53	6.72	6.89
4. S.D. of production	0.0	0.0	0.0	0.0	0.0	0.04
5. Exp. total cost (Rs. million)	866.3	909.5	972.1	1076.5	1135.6	1202.5
6. S.D. of total cost	11.9	31.9	45.3	43.9	46.1	43.9
item	1991-92	1992-93	1993-94	1994-95	1995-96	2000-01
1. No. of exploratory wells	28	31	34	37	40	42
2. Exp. discovery (mmt)	2.12	0.50	0.12	0.01	0.0	0.0
3. Exp. Total production (mmt)	7.08	7.35	7.66	7.99	0.0	0.0
4. S.D. of production	0.09	0.17	0.23	0.30	0.0	0.0
5. Exp. total cost (Rs. million)	1247.9	1317.1	1360.3	1346.2	0.0	0.0
6. S.D. of total cost	36.6	47.6	56.9	70.0	0.0	0.0
item	1985-86	1986-87	1987-88	1988-89	1989-90	1990-91
		N = 60				
1. No. of exploratory wells	16	17	18	20	22	25
2. Exp. discovery (mmt)	0.0	37.15	38.14	33.87	35.00	33.27
3. Exp. total production (mmt)	6.28	6.24	6.22	6.53	6.72	6.89
4. S.D. of production	0.0	0.0	0.0	0.0	0.0	0.0
5. Exp. total cost (Rs. million)	867.0	924.2	996.5	1115.4	1199.5	1292.0
6. S.D. of total cost	15.5	50.3	64.6	80.1	82.8	85.0
item	1991-92	1992-93	1993-94	1994-95	1995-96	2001-01
1. No. of exploratory wells	28	31	34	37	40	42
2. Exp. discovery (mmt)	30.35	27.29	21.06	16.11	9.33	0.0
3. Exp. total production (mmt)	7.11	7.42	7.82	8.36	10.53	10.71
4. S.D. of production	0.16	0.28	0.43	0.57	0.73	1.21
5. Exp. total cost (Rs. million)	1273.0	1497.4	1616.8	1736.4	1817.8	1878.9
6. S.D. of total cost	102.5	113.3	127.7	137.2	131.8	120.1

There are some features of these tables which deserve further explanation. Observe that standard deviations in first few columns are zero. It only indicates that all the discovered fields has a fixed profile for exploitation as described in the paper. Uncertainty and hence variation comes only for fields to be discovered. But the fields which would be discovered in the subsequent years have some lags for productions. After one starts an exploration, it needs time to dig wells and thus some lag is introduced before discovery. Hence for the first year the expected discovery is zero. Given the exploratory activity, after some time all the fields will be discovered and hence no more discovery is possible afterwards. When we talk in terms of expected discovery, the size of the expected discovery becomes too small to count after some time. Similar comments hold for production also.

On the whole, the results presented in these tables clearly bring out the implication of rising costs of further production effort in the basin. Since the basin is already fairly explored, a relatively large marginal effort is necessary for additional discovery, and hence the marginal cost of production is likely to follow a rising trend over time. As it is intuitively clear, the expected discovery is more in the case $N = 60$ than in the case $N = 40$ and the basin takes some more time, then $N = 40$, to be fully discovered.

5. **Some Unsolved Problems**

There are a few interesting unsolved statistical problems which have come up in course of this work.

(a) The non-stochastic econometric approach mentioned in Section 2.2.2 postulates a non-linear regression model of the form

$$R_E = K - a(b - C_E)^r, \quad 0 < C_E < b \qquad \ldots (5.1)$$

between R_E = cumulative reserve established and C_E = the corresponding cumulative cost of production.

The relation (5.1) or something similar should follow naturally as a long term average relation based on the stochastic model for discovery and cost. For this purpose it is necessary to study the asymptotic behaviour of the remaining reserve

$$E(Y_{n+1} + \ldots + Y_N \mid y_1 \ldots, y_n)$$

and the corresponding expected cost

$$E(T_{n+1} + T_{n+2} + \ldots + T_N \mid y_1, y_2, \ldots, y_n) \; as \; N \to \infty.$$

Presumably suitable average values of $y_1 \ldots y_n$ may have to be chosen in this exercise and the Gamma distribution may have to be truncated to use appropriate central limit Theorem for Y_i's (vide Rosen (1972), Gordon (1983a, 1983b) etc.), other assumptions as discussed in the present article remaining unaltered. It is conjectured that $w(y) \propto y^{\delta}$ is crucial in arriving at a long term average relation like (5.1).

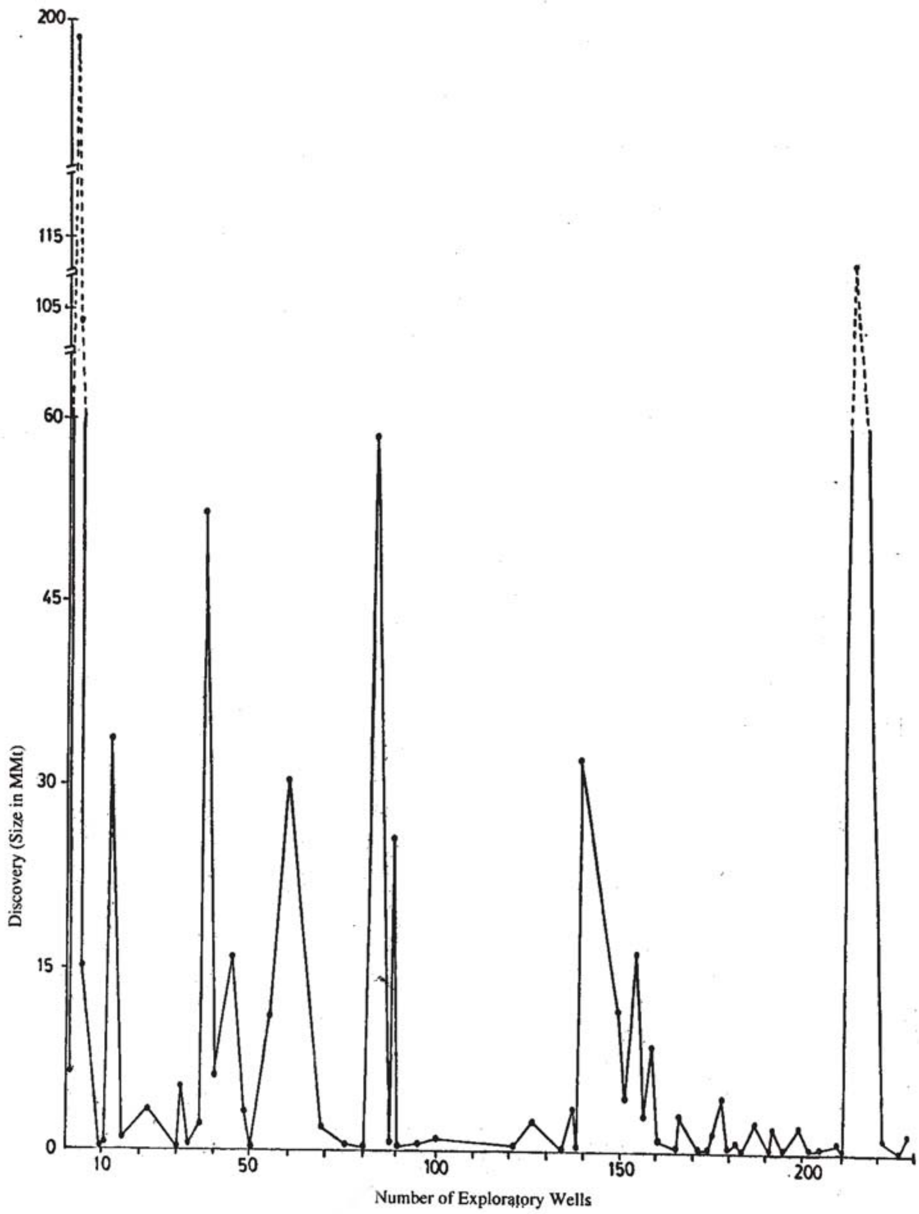
200
115
105
60
45
30
15
0
Discovery (Size in MMt)
10
50
100
150
200
Number of Exploratory Wells

Figure 5

(b) The work of Bickel, Nair and Wang (1992) seems to suggest that an estimate of N in our model based only on y's and having good asymptotic properties should be available. It would be interesting to study the properties of the maximum likelihood estimate. Incidentally, if one takes the unknown remaining y's as the parameters the MLE for R and N are pathological. They are equal to 0 and n respectively.

The data relating to discoveries and their sizes for the basin considered in Section 4 are presented below.

Table 5.1 : TOTAL AND AVERAGE RESERVES ESTABLISHED THROUGH 63 DISCOVERIES FOR DIFFERENT VALUES OF W_n

W_n	f	total reserve (mmt)	average size (mmt)
0	18	443.1159	24.62
1	8	51.1560	6.39
2	13	209.8470	16.14
3	5	7.6880	1.54
4	9	39.0815	4.34
5	2	0.9700	0.49
6	3	35.7600	11.92
7	3	1.2970	0.43
9	1	11.8400	11.84
21	1	0.4000	0.40
Total	63	801.1584	

mmt : million metric ton.

Acknowledgement. The I.S.I. interdisciplinary team in charge for the work on ONGC. consisted of Robin Mukherjee, Dipankar Coondoo and T.S. Arthanari apart from the present authors. J.K. Ghosh acted as a consultant for the team.

All members of the team contributed to the understanding and solution of the present problem. Their help at various stages is gratefully acknowledged.

The authors also acknowledge ONGC. for giving permission to use data necessary for the study in this paper.

Constructive comments of four referees have led to substantial improvement in presentation.

References

ANDREATTA, G. AND KAUFMAN G.M. (1985). Estimation of finite population properties when sampling is without replacement and proportional to magnitudes, *MIT Working Paper*, 82-027 WP, Cambridge, M.A.

ARPS, J.J. AND ROBERTS, T.G. (1958). Economics of drilling for getaceous oil on east flank of denver - julesburg, basin, *AAPG Bulletin*, 42, 2549-2566.

BAROUCH, E. AND KAUFMAN, G.M. (1976a). Probabilistic modelling of oil and gas discovery in energy : *Mathematics and Modelling*, ed. F. Robert Philadelphia, SIAM.

BAROUCH, E. AND KAUFMAN, G.M. (1976b). Oil and gas discovery modelled as sampling proportional to random size, *Working Paper*, WP 888-76, Sloan School of Management, KJT.

BALCER, Y., KAUFMAN, G.M. AND KRUST D. (1975). A probabilistic model of oil and gas discovery. In Studies of Geology No. 1 - Methods of Estimating Volume of Undiscovered Oil and Gas Resources, *AAPG* Bulletin, 113-42.

BERGER JAMES D. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second Edition, Springer Verlag, New York, Berlin, Heidelberg, Tokyo.

BICKEL, P.J., NAIR V.N., WANG P.C.C. (1992). Nonparametric inferences under biased sampling from a finite population, *Ann. Statist.*, Volume 20, 853-878.

BLOOMFIELD, P., DEFFEYES, K.S., WATSON G.S., BENJAMINI, Y., AND STINE, R.A. (1979). Volume and area of oil fields and their impact on order of discovery. *Report of the Resource Estimation and Validation Project*, Princeton University, Depts. of Statistics and Geology.

DREW, L.G., SCHUENEMEYER, J.H., ROOT, D.H. (1980). An application to the denuer basin, in Petroleum-Resource and Discovery Rate Forecasting in Partially Explored Regions, *U.S. Geological Survey Professional Paper*, 1138 - Parts A.B.C., U.S. Dept. of the Interior.

EPPLE, D. AND HANSEN, L.P. (1979). An econometric model of U.S. petroleum supply work optimal endogeneous depletion, *Carnegie-Mellon Univ., Working Paper.*

ERRICKSON, E., AND SPANN, R. (1971). Supply responses in the regulated industry : the case of natural gas, *Bell J. of EC. and Management Science*, 2, 94-121.

GORDON, L. (1983a). Successive sampling in large finite populations, *Ann. Statist.*, 702-706.

– – –– (1983b). Estimation for large successive samples with unknown inclusion probabilities, *Energy Information*, U.S. DOE, Washington D.C.

I.S.I. - ONGC COLLABORATIVE RESEARCH PROJECT (1990). Estimation of discovery and production costs of hydrocarbon with some applications in India data : *Unpublished Report Submitted to ONGC of India.*

KAUFMAN, G.M., BALCER, Y., AND KRUYT, D. (1975). A probabilistic model of oil and gas discovery, in J.D. Haun (ed.), Methods of Estimating the Volume of Undiscovered Oil and Gas Resources : *AAPG Studies in Geology*, 1, 113-142.

KHAZOOM, J.D. (1971). The FPC staff's econometric model of natural gas supply in the U.S., *Bell J. of EC. and Management Science*, 2, 51-93.

LEE, P.J. AND WANG, P.C.C. (1983a). Probabilistic formulation of a method for the evaluation of petroleum resources, *Mathematical Geology*, 15, 163-181.

– – –– (1983b). Conditional analysis for petroleum resource evaluations, *Mathematical Geology*, 15, 349-361.

MAITI, P. AND PAL, M. (1993). The process of exploration and exploitation of hydrocarbon - a stochastic modelling, *Presented at P.C. Mahalanobis Birth Centenary Celebrations Symposium on Sample Surveys : Theory and Methods*, Calcutta, 15-17, 1993.

MC CRAY, A.W. (1975). *Petroleum Evaluation and Economic Decisions*, Prentice Hall, Englewood Clifs, New Jersey.

NAIR, V.J. AND WANG, P.C.C., (1989). Maximum likelihood estimation under a successive sampling discovery model, *Technometrics 3.1*, 423-436.

NEWENDROP, P.D. (1975). Decision analysis for petroleum exploration, *Petroleum Publ. Co.*, Tulas, Oklahama.

O'CARROLL, F.M. AND J.L. SMITH (1980). Probabilistic methods for estimating undiscovered petroleum resources, *Advances in the Economics of Energy and Resources*, 3, 31-63.

RAMSEY, J. B., (1981). The economics of oil exploration : a probabilye of - ruin approach, *Energy Economics*, 14-30.

ROSEN, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement, I and II, *Ann. Math. Statist.*, 43, 373-397; 748-776.

SMITH, J.L. AND WARD, G.L. (1981). ML estimates of the size distribution of north sea oil fields, *Mathematical Geology*, 13, 399-413.

WEST MIKE (1992). Discovery, sampling and selection models, *Duke University, Discussion Paper No. 92-A06.*

J.K. GHOSH, P. MAITI, A.C. MOKHOPADHYAY, M. PAL
INDIAN STATISTICAL INSTITUTE
203 BARRACKPORE TRUNK ROAD
CALCUTTA 700 035
INDIA
e-mail : pulakesh@isical.ac.in