Distributions with Beta properties as a model for the parent size distribution of grown oil and gas pools.

Jostein Lillestøl[1], Richard Sinding-Larsen[2], Kenneth C. Hood[3]

Abstract Economic analyses are often highly sensitive to accumulation size, thereby influencing when and at what cost a potential resource will become available. For established plays, it may be expedient to base predictions of future field sizes on previous discoveries. In most cases, the historical size distribution is inappropriate to apply to the remaining potential within an established play because the largest accumulations tend to be discovered early in the exploration process. This intentional sampling bias must be accounted for when projecting sizes of future discoveries, as does the tendency to underreport the ultimate recoverable volume of hydrocarbons early in the life of a field. Exploration can be represented as a sampling process that both is size biased (creamed) and truncated. One can in cases where this representation is appropriate obtain an unbiased estimate of the resource base population (parent population) parameters by correcting for the bias (un-creaming).This can be achieved by explicitly modeling the relation between the sample parameters and the population parameters. Field size data from the Gulf of Mexico shelf were used to explore the relationship between discovered and future field sizes. Using Beta distributions and distributions with the same uncreaming property as assumed parent populations, it is possible to simulate discovery sequences that closely match the historical development of the basin. A discovery sequence can be approximated using lognormal distributions across multiple stages of exploration and over a wide distribution of discovery sizes. These observations suggest that one cannot impute the shape of the underlying parent distribution from the size distribution of past discoveries. An advantage, however of using a Beta distributed parent population is that the creaming bias is represented by a single parameter that can be estimated from the discovery sequence and subsequently used for an unbiased estimation of the parent population that potentially can incorporate a larger number of small fields than the lognormal distribution, and thus may significantly impact play economics.

1 Department of Finance and Management Science, Norwegian School of Economics (NHH), Helleveien 30, 5045 Bergen, Norway, Jostein.Lillestol@nhh.no

2 Department of Geology and Mineral Resources Engineering, Norwegian University of Science and Technology (NTNU), Sem Sælands vei 1, N-7491 Trondheim, Norway, Richard.Sinding-Larsen@ntnu.no

3 Exxon Exploration Company, 440 Benmar, Houston, TX 77060, USA, Kenneth.C.Hood@ExxonMobil.com

Introduction

A histogram of the number of accumulations discovered in an assessment region is called an accumulation-size distribution. It shape reflects the parent size distribution of hydrocarbon accumulations (resource base) that exist in the assessment region, the efficiency of the discovery process, and economic constraints that limit the development of extremely small accumulations. Traditionally, accumulation-size distributions are approximately lognormal in form, with a pronounced tail extending to the larger accumulation sizes on the right. In a maturely explored assessment region, the right tail of the accumulation-size distribution of discoveries will closely correspond to the shape of the underlying distribution of accumulations originally in place (the resource base) because creaming has resulted in the discovery of almost all the largest accumulations. The lognormal model must accordingly be truncated to be able to represent the finite volume of even the largest hydrocarbon accumulation. The shape of the left tail, however, reflects both the creaming effect as well as economic truncation. Increases in hydrocarbon prices may shift the point of economic truncation to the left, so previously uneconomic discoveries may be placed into production. Most of the remaining undiscovered economic potential of mature assessment regions may lie in economically sub-marginal accumulations. The distribution of possible sizes of undiscovered accumulations that exist in a proven play is conditional upon the underlying finite parent population of accumulation sizes and the creaming effect of the discovery process that can be regarded as the result of biased sampling without replacement of the accumulations in the play.

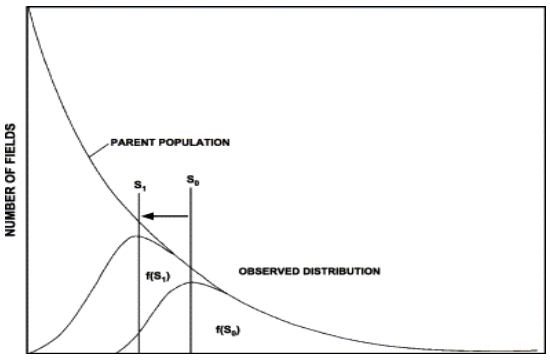


Fig. 1 Left shifting of accumulation size distribution f(S0) to f(S1) and of cost truncation point S0 to S1 on the parent population, due to price rise. (Drew,L.J.,1997).

There are several ways of estimating the parent accumulation size distribution. One common way, when the play under consideration has data on the sizes of the discovered accumulations, is by defining the accumulation size distribution from the frequency distribution of the sizes of the discovered accumulations in the play. This is feasible only when the sample size of accumulation sizes is sufficient enough, so that the shape of the frequency distribution can be taken to 'mimic' at least 'some part' of the shape of the parent population. Many critical issues are related to the empirical frequency distributions and the problems faced in defining the shape of the parent populations. In particular, left truncation in the empirical distribution related to the economics of hydrocarbon exploration and left shifting of the mode of the distribution due to the creaming effect of the discovery process (fig.1). After this introductory statements let us clarify some

definitions and concepts. The hydrocarbon resource endowment refers to the natural occurrence of hydrocarbon accumulations within a given assessment volume and is conceptually purely physical and is not dependent upon technology and economics. Resources refer to hydrocarbons contained in accumulations, which, if they were discovered, could be technically and economically produced today or in the near future. Hence resources are a function of the original endowment, economics and technology. The resource base is intermediate to the endowment and reserves and refers to the totality of hydrocarbons in accumulations equal or larger than the smallest accumulation size (Resource Base Minimum (RBM) included in the assessment.

Exploration can be represented as a sampling process that both is size biased (creamed) and truncated. One can in cases where this representation is appropriate obtain an unbiased estimate of the resource base population (parent population) parameters by correcting for the bias (un-creaming).This can be achieved by explicitly modeling the relation between the sample parameters and the population parameters considering the lower truncation (a) at the resource base minimum (RBM) size and the upper truncation (b) at the $a \leq x \leq b$ resource base maximum.

Traditionally lognormal, [1], or Pareto models [3 have been chosen to represent the accumulation size distributions. Beta distributions have been used for modeling input parameters in reserve and resource estimation [4]. In this extended abstract an alternative use of the Beta distribution is presented that have many desirable properties useful for predicting undiscovered sizes in a play.

Let x be the size of a hydrocarbon accumulation. Suppose that X = Beta (p,q,a,b) represent the four-parameter beta pdf for x as $f(x,p,q,a,b)$ where the parameter p controls the lower tail and the parameter q the upper tail . Considering that the probability for x to be discovered is proportional to x^d provided that $a \leq x \leq b$ where a is the lower and b is the upper truncation size then the pdf for discoveries $f^{d}(x;p,q,a,b)$ is related to the pdf for the accumulation size of the resource base in the following way:

$$f^{d}(x;p,q,b) = \frac{x^{d} f(x;p,q,b)}{\int_{a}^{b} x^{d} f(x;p,q,b)dx}$$

The implication is that if the largest accumulations that represent all accumulation sizes larger than a given threshold that are present in the assessment area have been discovered then an estimate of (q) by these discoveries should be equal to the (q) of the un-creamed parent resource base distribution. We are therefore, when it is reasonable to assume that the assessment area is sufficiently explored so that all accumulations above a known size x* have been discovered, left with the estimation of p and d. The existence of both high and small sizes in the start of the discovery sequence may reflect a contaminated distribution with a d= 0 interacting with a larger creaming factor for d. Estimating this contamination jointly with the estimation of p and d is proposed to represent a desirable strategy for approaching a permissible model for the parent distribution.

The abstract is organized as follows: In Section 2 we introduce biased sampling and its relation to discovery creaming; in Section 3 we discuss GOM field size distribution. In Section 4 we show the results of simulating discoveries analogue to the GOM fields. Finally, in Section 5 we discuss results and present conclusions.

2 Biased sampling: Discovery creaming

Given undiscovered oil accumulations in the ground with magnitudes that, varies according to an assumed distribution. Over time some of the accumulations are drilled and the resources found are recorded, and used to confirm or establish the characteristics of this distribution, say mean, median and mode. However, the sampling is scarce, and the most promising accumulations are likely to be drilled first, based on some geologic indicators ("creaming"). This will give a biased view of the underlying distribution. The distribution will typically be skewed with a long right tail, and the popular lognormal distribution is frequently used to represent the variation. In order to analyze the consequences of creaming, it may as proposed by [6] be illuminating to consider PPS- sampling, i.e. sampling with probabilities proportional to (expected) size. Distributions with an unrestricted right tail, like the lognormal, do not allow this, and the problem will instead be studied within the class of Beta-distribution. Although this is a distribution over the interval [0,1], it can be scaled to any interval [a, b}, and it accommodates skew distributions, which in small and moderate samples turn out to be indistinguishable from the log-normal as shown in figure 2.

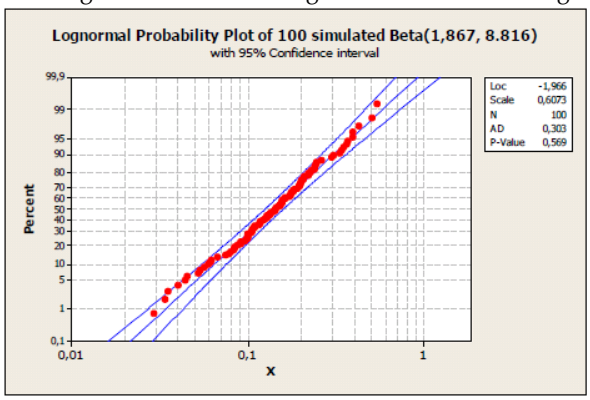


Fig. 2 Lognormal Probability Plot of 100 simulated Beta(1,867,8.816) with 95% Confidence interval

The Beta(p,q) density f(x) is proportional to $x^{p-1}(1-x)^{q-1} \quad 0 < x < 1$
The parameters p>0 and q>0 determines the shape of the distribution.

If X=x is sampled with probability proportional to x, then the sampled X becomes Beta (p+1,q) instead of Beta (p,q). Similarily if the sampling is with probabilities proportional to any power xd of x then sampling will appear as coming from a Beta (p+d,q) distribution instead of the true Beta (p,q). How much one should degrade the computed "p"=p+d in order to get at the "true" p depends on how efficient the explorers go after the large expected deposits. This may be decreasing as time goes by, so that it is natural to expect that d depends on the current p. If the true p is estimated by a fraction of p+d, say by f·(p+d) then as p decreases, so will d and we have a dependency that reflects the expected exploration efficiency.

3 Gulf of Mexico fields

Some empirical results from a real exploration process may reveal patterns that might turn useful for prognostic purposes. Field sizes collected by the Mineral Management Service (MMS) from a mature exploration area of the U.S. Gulf of Mexico (GOM) Shelf up to 2002 representing "Proved" reserves totaling N=982 observations are shown in the histogram in figure 3. We see an abundance of small sizes and a few large outlying ones. The sizes range from close to 0 to less than 800

MMbbls o.e., and were rescaled accordingly to [0,1]. The scaled sizes in the actual order revealed and recorded are given in the time series plot on the right in fig.3. We see a pattern that is not consistent with i.i.d. random sampling, but a pattern declining over time and some a large number of small sizes even in the beginning of the sequence. An obvious advantage of assuming a Beta distributed parent population is that the effect of creaming is summed up in the single parameter p and it is therefore interesting to explore the possibility to keep the model framework at this simple level. As basis for this we may look at some empirics from the data:

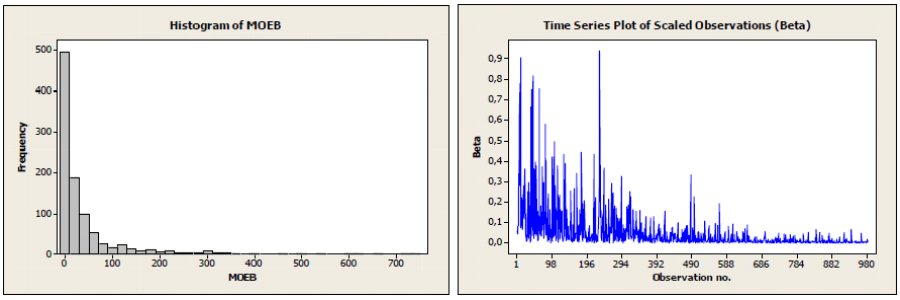


Fig. 2 Histogram and discovery sequence of GOM shelf fields MOEB (millions barrels oil equivalents).

Let us consider the range of GOM field sizes divided into four time segments Q1, Q2, Q3 and Q4 of about equal size, here with number of observations 245, 245, 245 and 247 respectively.

Table 1 Empirics for the first 4 time segments of the GOM field sizes basin.

	No. of obs.	Mean	Variance	"p" (p+d)	q
Q1	245	0.1319	0.0284	0.3992	2.6272
Q1+Q2	490	0.0894	0.0178	0.3198	3.3257
Q1+Q2+Q3	735	0.0649	0.0133	0.2321	3.3417
Q1+Q2+Q3+Q4	982	0.0504	0.0106	0.1775	3.3419

The computation of distribution parameters based on enlarged segments is shown in table 1. We see that the "p" decreases steadily and q increases slightly, and rapidly stabilizes after the first period, consistent with p mainly affecting the left tail and q the right tail. After the big fields are sampled, we do not expect much to happen in the right tail, but changes are still expected in the left tail. If we do the encompassing computation sequentially and plot the resulting p+d and q as function of the number of observations (omitting the first 25 values which are very erratic due to few observations) we get the plots in fig. 4. A decline in p+d with n, and an apparent asymptotic behavior in q can be observed. The tail behavior of p+d appears slightly convex, but not far from linear. As n goes to infinity we essentially get the sizes of the parent distribution, but we get them in an order not consistent with independent sampling (not even independent with probabilities proportional to size). A possible context is therefore: The geological processes which have distributed field sizes in the subsurface above a minimum size represents a random sample of N from a continuous parent distribution, here assumed a rescaled Beta. What really is observed is a discovery sequence in an order related to the size of each field. A way of exploring this context is to simulate randomly from Beta(p,q) and then order the observations according to size, or some principle related to magnitude, and then see how successive encompassing calculations reflect the observed (tail) behavior.

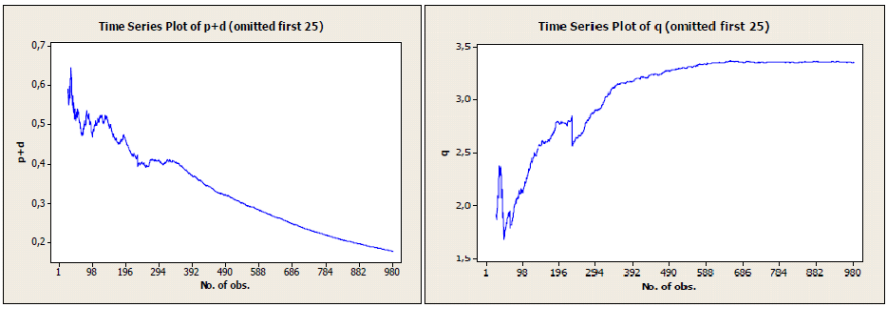


Fig. 2 Time series plot of p+d and q

4 Simulated Discoveries

1000 independent Beta(0.5,2.5) are generated to constitute a population from which the 1000 potential discoveries are selected without replacement, one at a time, with probabilities proportional to x^d among the remaining hydrocarbon accumulations.

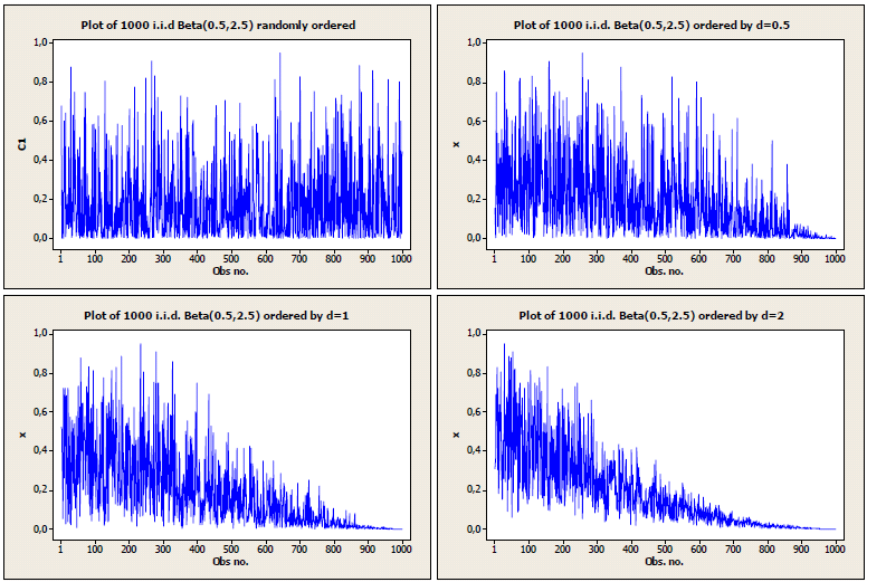


Fig. 1 Simulation plots for d=0 (top left), for d=0.5 (top right), d=1 (bottom left) and d=2 (bottom right):

The result shown in figure 5 have features common with the corresponding plot for the real GOM data in fig. 3. However, there are some differences. The existence of both high and small sizes for some time is more consistent d=1, but in this plot tails off too slowly. On the other hand d=2 tails off more consistent with the real case, but lacks the small sizes in the beginning. For the real GOM data we have both high and low ones for some time. This may reflect special features of the exploration process, among them that seismic provide better control of the large fields, in the sense that it is not that easy to avoid the small discoveries as it is to find the large ones. Appearance of new information and practical reasons may also affect the process. A realistic model of this kind may therefore be represented by a mixture between d=0.5 and d=2 that would give a pattern close to the one observed, while a single creaming factor d=1 represents a compromise.

It is of interest to see how successive computations of the Beta-parameters differ in the four cases d=0, 0.5, 1, 2.

4.1 Discovery as a mixture of two exploration processes

Simulated drawings from the same population of 1000 has been generated from Beta(0.5,2.5) as in the previous simulation. The drawings are according to the same scheme, but each item is sampled according to d=0 or 1 with probabilities (1-r, r) for r=0.3, 0.5, 0.7. This results in the following plots shown in fig. 6. For the p- parameter we observe the same downward slope as for d=1 previously. For the q-parameter we have an upward slope for r=0.3, tailing off asymptotically for r=0.5, and still a peak late in the sample process for r=0.9.

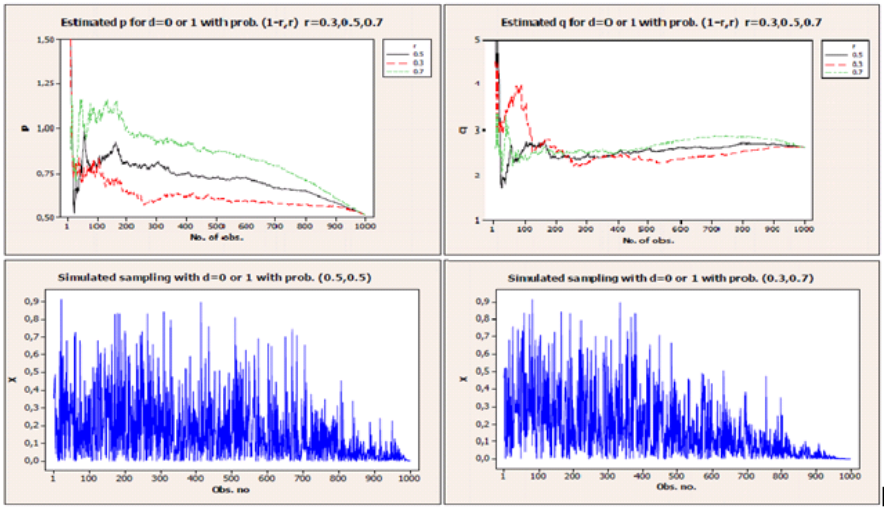


Fig. 6 Estimated p and q (top), discovery sequences (bottom)In the plots the first 10 calculated values are omitted, in order to get higher resolution that otherwise would have been spoiled by high and erratic values.

The pattern most consistent with the plots for the real GOM discovery process of is for r=0.5. An estimate for r can be derived from the exploration success rate and the creaming factor. The graphs of the simulated sampling process is given here for r=0.5 and r=0.7. We see a slight difference in how it tapers off at the end, linearly for r=0.5 and exponential for r=0.7. This feature is supported by repeated simulations. The main difference is that the real discovery sequence has both high and small sizes in the beginning, more uniformly up to a midway in the exploration process, where we have a major drop to uniformly small sizes and not the kind of tapering off as in the graphs below. The sampling process for the real GOM data in section 3 seems to be more consistent with a mixture between d=0 and d>1 (say 2) until the large sizes are almost depleted, from which time the remaining discovered sizes occur more randomly.

4.2 Time until largest discovery and population size

The context is as before, a population of N field sizes given as a random sample from a scaled Beta parent distribution. There are several ways of reasoning, and one possibility is to focus on the maximum size observed in the sampling process, and utilize that the expected time before discovering the largest field will be dependent on N. By observing the time of the maximum it should be possible to project N. This may work since within any size based sampling scheme, we are fairly sure that we observe the maximum in a reasonable time compared to the size of the population. It seems hard to develop analytic formulas, and we will resort to simulations in order to establish the relationship. Beta(0.5,2.5) is used as a model for the parent

superpopulation and successively subsidiary populations of size N= 10, 20, 30, …, 1000 are simulated. From each of these subpopulations hypothetical discoveries are drawn according to the mixture proportional to size scheme (d=0,d=1) with probabilities (0.5, 0.5) for r =0.0, 0.1, 0.2, …,0,9, 1.0. For each simulation the waiting time for the maximum discovery is we observed. The average $\overline{n}_{max}$ of the observed waiting times for the maximum was taken as an estimate of the expected waiting time calculated from 100 repetitions of each combination of parameters. The plot in fig. 7 of N versus $\overline{n}_{max}$ for the case d=1 and r=0.5, have a clear linear structure.

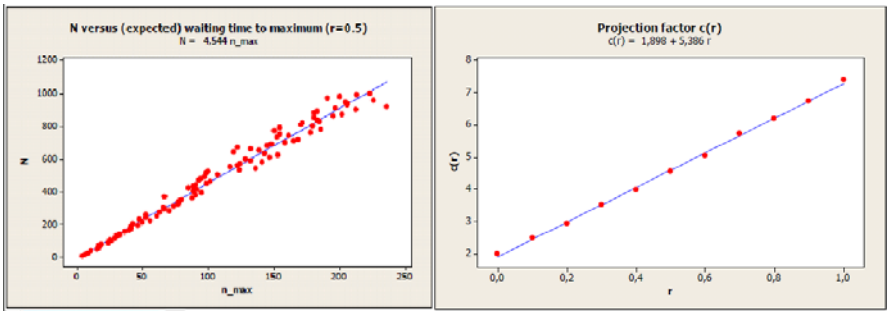
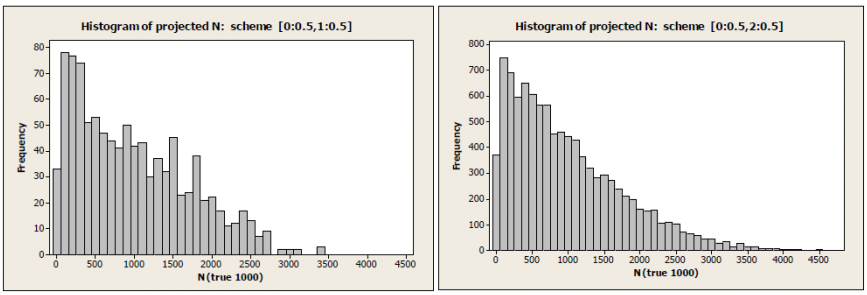


Fig. 7 N versus $\overline{n}_{max}$ (left) and c(r) versus r (right).

Simulation is then repeated for (0,d) with probabilities (1-r, r) for r =0.0, 0.1, 0.2, …,0,9, 1.0, in the case of d=1 and d=2, all showing a similar linear structure with c(0)=2 and increasing slopes. This suggests the projection formula $N=c\overline{n}_{max}$ where, for each r, we may compute c=c(r) from regressing N on $\overline{n}_{max}$ as shown in the right side of the plot in figure 7. The waiting time for the maximum for a given N may be far off the expectation line in a single application (fig.8). Consequently the projection of N, could be far off if an individual $\overline{n}_{max}$ value is used as a basis for projection unless constrained by auxiliary information.



Descriptive Statistics: Projected N

Variable	obs.	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Max
(0,1) N	10000	978	7,18	718	4,54	378	848	1450	3538
(0,2) N	10000	977	7,75	775	9,24	360	794	1432	4749

Fig. 8 Histogram of projected values for N

5 Simulations and conclusions

Simulations for a specific parent population, and assumed specific proportional to size mixture sampling scheme (0,d) for d=1,2has shown that the projection factor for estimating the population size N is heavily dependent on r the mixing probability. The effect of simultaneous varying r, p and q is under study. The simulation match fairly well with the real exploration sequence observed from the GOM shelf, and has by this permitted to explore some of the challenges of extrapolating discovery sequences. We may face a quite different discovery pattern in another exploration play, and will have to learn as we go along, and simultaneous being asked to make

projections before we really know how the exploration sequence behaves. However, there is some hope that before the play reaches half maturity sufficient insight is acquired to make an educated guess about the number and sizes of remaining discoveries. An advantage of using a Beta distributed parent population is that the creaming bias is represented by a single parameter that can be estimated from the discovery sequence and subsequently used for an unbiased estimation of the parent population that potentially can incorporate a larger number of small fields than the lognormal distribution, and thus may significantly impact play economics.

References

[1] L.J Drew Undiscovered Petroleum and Mineral Resources; Assessment and Controversy Plenum, New York (1997)

[2] P. J. Lee, Oil and gas pool size probability distributions:J-Shaped. lognormal or Pareto?: Geol. Survey Canada, Current Research, Part E, Paper 93–1E, p. 93–96. 1993

[3] Z.Chen and R. Sinding-Larsen. Estimating number and field size distribution in frontier sedimentary basins using a Pareto model Natural Resources Research, Volume 3, Number 2, Pages 91-95. 1994

[4] R. Olea. On the Use of the Beta Distribution in Probabilistic Resource Assessments *Natural resources research* Volume 20, Number 4, 377-388, Springer, 2011

[5] C .Kleiber. and S.Kotz. Statistical Size Distributions in Economics and Actuarial Sciences. *Hoboken, NJ: John Wiley and Sons*. 2003

[6] E.Barouch, G.M.Kaufman, Oil and gas discovery modelled as sampling proportional to random size. Cambridge, Mass: M.I.T. Alfred P. Sloan School of Management (http://hdl.handle.net/1721.1/48701) , 1967.

[7] G. M. Kaufman, , Finite population sampling methods for oil and gas resource estimation, in Rice, D. D., ed., Oiland Gas Assessment—Methods and Applications: Am. Assoc. Petroleum Geologists, Studies in Geology No. 21, p. 43-53.198

[8] K. C. Hood, L. M. Wenger, O. P. Gross, and S. C. Harrison AAPG Studies in Geology No. 48 / SEG Geophysical References Series No. 11, Chapter 2: Hydrocarbon Systems Analysis of the Northern Gulf of Mexico: Delineation of Hydrocarbon Migration Pathways Using Seeps and Seismic Imaging, Pages 25 – 40. 2002

[9] M.B.Gordy. A generalization of generalized beta distributions. Board of Governors of the Federal Reserve System. Finance and Economics Discussion Series. 1998