MIT Libraries | DSpace@MIT

MIT Open Access Articles

Has the Largest Field Been Discovered Yet? PETRIMES and GRASP 25 Years Later

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Kaufman, Gordon M., Ray Faith, and John H. Schuenemeyer. "Has the Largest Field Been Discovered Yet? PETRIMES and GRASP 25 Years Later." Mathematical Geosciences 48.8 (2016): 873–890.

As Published: http://dx.doi.org/10.1007/s11004-016-9652-z

Publisher: Springer Berlin Heidelberg

Persistent URL: http://hdl.handle.net/1721.1/106325

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike

Massachusetts Institute of Technology

Has the Largest Field Been Discovered Yet? PETRIMES and GRASP 25 Years Later

August 11, 2016

Received: _____ / Accepted: _____ / Published online: _____

Gordon M. Kaufman[1], Ray Faith[2], John H. Schuenemeyer[3]

[1] Corresponding author: Massachusetts Institute of Technology Sloan School of Management
address: 77 Massachusetts Avenue, Cambridge, MA
e-mail: gkaufman@mit.edu
phone: 617-253-2651
[2] Massachusetts Institute of Technology Sloan School of Management, Cambridge, MA
[3] Southwest Statistical Consulting, LLC, Cortez, CO.

Abstract Assessment of undiscovered oil and gas resources has been an important component of energy policy for the governments of the United States and Canada for many years. A pool-size-by-rank statistical procedure is a centerpiece of the Geological Survey of Canada's Petroleum Exploration and Resource Evaluation System (PETRIMES and of the U.S. Department of Interior's Geological Resource Assessment Program (GRASP). Both employ discovery process modeling to make inferences about the number of pools in a play and about parameters of the play's pool size distribution. The pool-size-by-rank procedure implemented in these two systems abandons a key primitive postulate on which modern discovery process models are based — sampling proportional to pool size and without replacement. This logical disjunction has consequences: the predictive distribution of number of pools remaining to be discovered and the predictive distribution of undiscovered pool sizes generated by use of pool-size-by-rank procedures differ substantially in shape, location and spread from predictive distributions that incorporate sampling proportional to size. Uncertainty about total undiscovered oil and gas in a play is diminished.

Keywords Oil and Gas Discovery Modeling, Successive Sampling, GRASP, PETRIMES, Pool Size by Rank, Sampling Proportional to Size

1 Introduction

The 1975 USGS nationwide oil and gas assessment of conventional resources pioneered the use of subjective probability methods for projecting undiscovered oil and gas in US petroleum provinces. Probabilistic projections in the 1975 study are subjective, although geological reasoning guided study participants in their assignment of probability distributions to in place and recoverable oil and gas in US petroleum basins (Miller et al. 1975). This exercise foreshadowed the next major shift in methods designed to appraise undiscovered petroleum: use of these resources discovery process models based on primitive assumptions about how they are discovered to generate probabilistic projections of remaining undiscovered oil and gas in a petroleum play as a function of discovery history. Both the Geological Survey of Canada (GSC) and the U.S. Bureau of Ocean and Energy Management (BOEM) adopted oil and gas discovery modeling to aid in assessment of undiscovered pools in U.S. and Canadian petroleum provinces. This paper identifies strengths and weaknesses of these models as devices for probabilistic projection of undiscovered oil and gas in petroleum plays. Section 1 presents a compact summary of essential features of oil and gas discovery process models used by the CSC and the BOEM. Section 2 briefly describes the history of use of discovery process modeling by these two agencies and lays the ground for Sect. 3, in which a logical disjunction in the pool-size-by-rank protocol is identified. Pool-size-by-rank is one of several procedures used by these agencies to appraise undiscovered conventional oil and gas. It was created in the 1980s to simplify computation of probabilistic prediction of oil and gas remaining to be discovered in a petroleum play. Methods for computing probability distributions of complicated uncertain quantities have grown explosively over the intervening thirty-five years. Markov Chain Monte Carlo (MCMC)

and importance sampling enable easy computation of quantities, which in the 1980s, appeared difficult to compute and time consuming (West 1994, 1996). PETRIMES and GRASP systems currently in use do not incorporate the latest statistical computation methods. In Sect. 4 we discuss enhancements to discovery process models.

Oil and gas discovery process models are designed to compute of predictive probability distributions for oil and gas remaining undiscovered in a petroleum play. Most account for the observation that, on average, large accumulations in a petroleum play are discovered before small accumulations. PETRIMES and GRASP discovery process models rest on two primitive postulates. The first is that that the empirical size distribution of accumulations in place in a petroleum play is generated by identically independently distributed (iid) sampling of accumulation sizes endowed with a probability density concentrated on $(0,\infty)$. Any reasonable choice of functional form is allowable. However, in the 1950s petroleum geologists with a statistical bent noticed that empirical frequency distributions of pool sizes in mature petroleum plays are approximately Lognormal (Blondel 1955; Allais 1957). Figure 1 is a Q-Q plot of log sizes of 2,509 Lloydminster play accumulations versus a Normal distribution computed by one of the authors using data provided by McCrossan et al. (1981). Pool sizes in this play vary by six orders of magnitude.

(Figure 1 Here)

The first key assumption is

Assumption I. Magnitudes $X_1,...,X_N$ of N in place accumulations in a petroleum play are generated by iid sampling $X_k \sim f(x|\theta),\ k=1,...,N$ of a density with domain $(0,\infty)$ indexed by a parameter $\theta \in \Theta$.

Empirical evidence such as Fig. 2 led many modelers to specify $f(x|\theta)$ as Lognormal indexed by parameter $\mu \in (-\infty, \infty)$ and $\sigma^2 \in (0, \infty)$ with domain $(0, \infty)$. A realization $x_1,...,x_N$ of $X_1,...,X_N$ creates a finite population of in place accumulation magnitudes on which discovery effort operates. As Fig. 2 shows, the second key idea is that large accumulations are on average discovered earlier in the discovery sequence than small accumulations.

(Figure 2 Here)

The outlier in the right hand graph is the Statfjord field. This and similar studies of discovery order by size in petroleum plays provide empirical support for Assumption II below.

Assumption II Given $\{x_1,...,x_N\}$ the probability of discovering $x_1,...x_N$ in the order $(x_1,...,x_n)$, $n \le N$ is

$$\prod_{k=1}^{n} \frac{x_k}{x_k + ... + x_{k+1} + ... + x_N} \quad . \tag{1.1}$$

In the finite population sampling literature, the acronym for this sampling scheme is SWORP; an alternative name is successive sampling.

Both PETRIMES and GRASP couple a super-population process for in place pool sizes to finite population successive sampling (Kaufman et al. 1975). Lee and Wang (1985) generalize Assumption II by introducing a discoverability parameter α

$$\prod_{k=1}^{n} \frac{x_k^{\alpha}}{x_k^{\alpha} + ... + x_{k+1}^{\alpha} + ... + x_N^{\alpha}} \quad , \alpha \in (-\infty, \infty) \tag{1.2}$$

When $\alpha = 0$ (Eq. (1.2)) becomes ordinary hypergeometric sampling, as $\alpha \to \infty$ accumulations are discovered exactly in order from largest to smallest and as $\alpha \to -\infty$ accumulations are discovered in order from smallest to largest. Combining I and II, the joint probability that N pool sizes $x_1 \in dx_1,...,x_N \in dx_N$ are in place and that successive sampling yields an ordered sequence $(x_1 \in dx_1,...,x_n \in dx_n)$) of pool sizes is

$$n!\binom{N}{n} \prod_{k=1}^{n} \frac{x_k^{\alpha}}{x_k^{\alpha} + ... + x_N^{\alpha}} \times \prod_{j=1}^{N} f(x_j|\theta)dx_j \quad .$$

(1.3)

If $f_{Z_n}(\lambda)$ is defined to be the density of a sum of n mutually independent exponential random variables with means $1/b_j(\alpha)$, $b_j(\alpha) = x_j^{\alpha} + ... + x_N^{\alpha}$, $j = 1,..,n$, the probability of observing a discovery sequence $(x_1 \in dx_1,..., x_n \in dx_n)$ and realizing undiscovered pool sizes $(x_{n+1} \in dx_1,...,x_N \in dx_N)$ is

$$n!\binom{N}{n} \times \prod_{j=1}^{n} f(x_j|\theta)dx_j \int_0^{\infty} f_{Z_n}(\lambda) \times [\prod_{k=n+1}^{N} e^{-\lambda x_k^{\alpha}} f(x_k|\theta)dx_k]d\lambda. \qquad (1.4)$$

so the joint density of $X_{n+1},...,X_N$ conditional on a discovery record consisting of an ordered sequence $(x_1 \in dx_1,...,x_n \in dx_n)$ of discovery sizes is

$$\frac{1}{C(\mathbf{s}_n, \theta, \alpha)} \int_0^{\infty} \prod_{k=n+1}^{N} [e^{-\lambda x_k^{\alpha}} f(x_k|\theta)\, dx_k]dF_{Z_n}(\lambda). \qquad (1.5)$$

Here F_{Z_n} is the cumulative distribution of a data dependent (discovered pool sizes) sum of independent but not identically distributed exponential random variables. Defining

$$L(\lambda \mid \theta, \alpha) = \int_0^\infty e^{-\lambda x^\alpha} f(x \mid \theta)\, dx, \tag{1.6}$$

$$C(\mathbf{s}_n, \theta, \alpha) = \int_0^\infty f_{Z_n}(\lambda) \times L^{N-n}(\lambda \mid \theta, \alpha)\, d\lambda\,. \tag{1.7}$$

The distribution of the data dependent sum Z_n plays a crucial role. Because of very large cancellations, numerical integration of Eq. (1.5) requires extreme accuracy. While feasible, this involves subtle programming and much calculation time. When this model was first proposed in the mid-1970s computation of one value of Eq. (1.6) to sixteen digits accuracy took four hours on a dedicated DEC 10! MLE done by steepest descent in a uniform asymptotic regime requires analysis of coalescing saddle points and computation of Airy functions (Barouch and Kaufman 1976). Viewed differently, this computational barrier becomes a computational blessing! West was the first to show that, in a Bayesian setting, acceptance-rejection sampling can be deployed along with MCMC to compute Eq. (1.5) and a variety of predictive distributions efficiently (West 1994, 1996).

For fixed parameter values N, θ and α, the predictive density of $X_{n+1},\ldots,X_N$ given $(x_1 \in dx_1,..,x_n \in dx_n)$ is a probability mixture of conditionally mutually independent random variables. Define $e^{-\lambda x^\alpha} f(x \mid \theta)\, dx_k = g(x \mid \lambda)$. Then the predictive density is

$$\frac{1}{C(s_n, \theta, \alpha)} \int_0^\infty \prod_{k=n+1}^{N} g(x_k \mid \lambda)\, dF_{Z_n} = \frac{1}{C(s_n, \theta, \alpha)} E_\lambda\{ \prod_{k=n+1}^{N} g(x_k \mid \tilde{\lambda})\} \tag{1.8}$$

making it evident that $X_{n+1},\ldots, X_N$ given $x_1 \in dx_1,\ldots, x_n \in dx_n$ possess identical marginal distributions and are symmetric with distribution independent of labeling.

If the number of discovered pools *n* is less than *N* a principal objective is to compute properties of the set of *N-n* undiscovered pool sizes based on their joint probability distribution

posterior to observation of a set of pool sizes in order of discovery. Given a discovery record, the probability distribution of a measurable function $g(X_{n+1},..,X_N)$ of $X_{n+1},...,X_N$ can be numerically approximated to a high order of accuracy by methods West advocates. This is true for a large variety of specializations: size or magnitude can be defined to be a function of a priori uncertain quantities such as rock volume, area, porosity, water saturation. For example, prospects can be distinguished from pools and the distinction modeled probabilistically, parameters controlling the discoverability of pools can be introduced, and pool sizes can be adjusted for reserve growth over time. Projections produced by pool-size-by-rank substantively differ from those produced by the sampling proportional to size model just described.

2 PETRIMES and GRASP Background

The Institute of Sedimentary and Petroleum Geology, Geological Survey of Canada, created the Petroleum Exploration and Resource Evaluation (PETRIMES) system in the 1980s to provide scientifically sound procedures for projecting undiscovered oil and gas in petroleum plays based on a discovery record. The US Department of the Interior's Minerals Management Service (now BOEM) built its version of PETRIMES called Geological Resource Assessment Program (GRASP) soon after. Discovery process models proposed in the 1970s are at the core of these systems. Walter Stromquist's 1998 review is an excellent summary supplemented with a detailed mathematical explanation of how PETRIMES works. An alternative summary of GRASP modeling methods and issues appears in OCS Report MMS 99-0034 (Lore et al. 1999). PETRIMES allows both a fully subjective approach to projection of undiscovered oil and gas and a discovery process model approach. The subjective approach is used in frontier basins with

little or no exploration history. Assessors provide subjective assessments of probability distributions of key geologic attributes. When a solid exploration history is available, discovery process modeling is appropriate. Both PETRIMES and GRASP are structured to employ Lee and Wang's pool-size-by-rank protocol for projection of undiscovered oil and gas in a petroleum play (1986). Pool-size-by-rank in particular was created to reduce computational complexity and to allow geologists to manipulate key parameter estimates produced by discovery process models imbedded in PETRIMES and GRASP so that final output appears reasonable.

In unpublished 2004 and 2011 reports the American Association of Petroleum Geologists Committee on Resource Evaluation summarized their reviews of hydrocarbon assessment methodology employed by the BOEM. In both reports they expressed concern that, in their judgment, GRASP's discovery history methodology under-represented uncertainty of the number and sizes of undiscovered conventional oil and gas fields in the Gulf of Mexico. The 2011 report states that the BOEM should not use statistical methods exclusively to address this methodological problem in favor of some type of subjective assessment method designed to increase uncertainty about both sizes and numbers of undiscovered fields. A quality statistical method is one that incorporates first principles, data when available and subjective judgment in a rigorous and reproducible manner. Unfortunately, the approach suggested in the 2011 report dodges a fundamental scientific issue: what constitutes an acceptable model for making inferences about undiscovered oil and gas in a mature petroleum play? A purely subjective approach to assessment raises its own problems. How thoroughly have those geologists been trained in probability assessment? In particular, have these geologists been calibrated to avoid assessment bias? Has the method employed to combine expert judgments been subjected to controlled experimentation designed to avoid a variety of pitfalls that commonly arise? The best

guarantee of both coherence and eliminating assessment bias is a yet to be invented omnibus method.

3 Pool Size by Rank

P.J. Lee's authoritative book entitled Statistical Methods of Estimating Petroleum Resources (2008) provides a detailed description of how PETRIMES is used to answer the questions posed above. His death is a true loss, but fortunately he was able to complete this compendium. He describes how he and Wang derived pool-size-by-rank distributions and how to use them. They begin by specifying either a fixed value for the number *N* of pools in a play or an a priori distribution for *N*. Next they use the record of pool sizes in order of discovery and the PETRIMES discovery process model to compute point estimates of the super-population distribution assumed to generate pool sizes. Finally, they compute a rank distribution for pool sizes assuming that super-population distributions are fixed at reasonable values. If, for example, the in place pool size distribution is Lognormal with parameters μ and σ^2, their procedure starts with computation of maximum likelihood estimates (MLEs) of μ and σ^2 given *N* and the discoverability parameter α. Then μ, σ^2 and *N* are manipulated to provide sensible fits to properties of order statistics of undiscovered pool sizes. Once the number of pools and super-population (field or pool size) distribution parameters are fixed, values of undiscovered pools are assumed to be mutually independent and identically distributed from the pool or field size distribution. This leads to projections of undiscovered pool sizes that can differ substantially

from predictive probability distributions conditioned on parameter estimates derived from a discovery process model likelihood function—a departure from established methods of statistical inference and a logical fork in the road.

If the process generating discoveries is modeled as successive sampling coupled with a super-population process for pool sizes, then conditional on observation of a suite of discovered pool sizes, undiscovered pool sizes are not mutually independent. (See Appendix A for a proof.) To be probabilistically coherent, properties of order statistics for undiscovered pool sizes should be based on the joint distribution (Eq. (1.5)) of undiscovered pool sizes given the discovery record. Using this joint distribution to calculate pool-sizes-by-rank can yield projections significantly different from those produced by the current GRASP pool-size-by-rank procedure.

Another issue afflicts pool-size-by-rank: probabilistic judgments about undiscovered pool sizes in a play posterior to observation of a suite of discoveries in it should be processed differently from judgments made prior to observation of these discoveries and that both types of judgments must be used in accordance with Bayes' Theorem. A bedrock principle governing the order of computation is:

A priori judgments about model parameters $\Rightarrow$ Observed discoveries $\Rightarrow$

Likelihood function $\Rightarrow$ Posterior distribution of model parameters $\Rightarrow$

Predictive joint distribution of undiscovered pool sizes

In Chapter Three (pages 45 to 47) of his monograph, P.J. Lee provides a detailed description of PETRIMES and in GRASP pool-size-by-rank procedures. On page fifty he says that it is based on the assumption that, given fixed super-population parameter estimates, rank statistics for the set of discovered pool sizes can be computed as if observations (discovery sizes) are drawn by iid sampling of the super-population distribution. Lee and Wang's pool-size-by-rank procedures

are based on this independence assumption (1983a, 1985, 1986). PETRIMES and GRASP codify it. Figure 3 is a pictorial summary of pool-size-by-rank projections for the Rimbey-Meadowlark play in Alberta, as calculated by Lee and Wang (1985, Fig. 8).

(Figure 3 Here)

3.1 The Discoverability Parameter

The discoverability parameter plays an important role in calculation of super-population parameter estimators but does not appear at all in GRASP rank r calculations. Given the discovered pool sizes

in a play, a subjective appraisal of the largest pool remaining to be discovered depends on the assessor's judgments about discovery process efficiency. If discovery is very efficient the largest pools are highly likely to be discovered early in the sequence of discoveries and the largest pool remaining to be discovered is likely to be small. On the other hand, as discovery becomes increasingly inefficient, the probability distribution of the size of the largest undiscovered pool approaches that of the largest order statistic for independent identically distributed random variables, each equipped with the super-population size distribution. This is longhand for saying that the distribution of the largest pool remaining to be discovered depends heavily on the assessor's judgments about the value assigned to a discovery process model's discoverability parameter (Lee and Wang 1985; West 1996).

P.J. Lee, a principal architect of PETRIMES, (Lee and Wang 19983a, 1983b; Lee and Tzeng 1993) and his co-workers understood the importance of conditioning order statistics calculations on the discovery record (Lee and Wang 1985). Nevertheless, current PETRIMES and GRASP pool-size-by-rank procedures are based on the assumption that the discoverability parameter is

zero, equivalent to assuming that undiscovered pool sizes are mutually independent and identically distributed with distribution that of the discovery process model super-population process—this, in spite of the fact that these systems both contain modules that compute maximum likelihood estimates (MLEs) of pool size parameters, the number N of pools in the play and the discoverability parameter. The Compute the Pool Size Rank (PSRK) module in the GRASP system does this. When there is a substantial discovery record, the assumption that the discoverability parameter is zero—when in fact it is not—can lead to large differences between the distribution of the largest remaining undiscovered pool size produced by the discovery process model and that produced by the pool-size-by-rank procedure.

Walter Stromquist's review of PETRIMES and GRASP is the most authoritative explanation of the underlying mathematics done to date. He provides a precise description of assumptions and calculations in all PETRIMES modules as of 1998. Stromquist's detailed mathematical derivation of the rank-r density function in his review of the PSRK model confirms that ranks are computed assuming that underlying random variables are iid from a super-population distribution (not necessarily Lognormal). Stromquist further challenges the assumption that pool sizes are generated by iid sampling from a single probability distribution for pool sizes from a different vantage point, noting that the PSRK module is absolutely reliant on a pool size model in which the number of pools is a random variable. Pool sizes are drawn independently from a super-population distribution and are also independent of the number of pools. Stromquist views this as reasonable because, in his opinion, there is no obvious alternative. To demonstrate that this assumption is not always appropriate, Stromquist visualizes two extreme scenarios: a play contains only a few very large pools or, in contrast, it contains many small pools. If these are the only possible cases, the number of pools in the play cannot be independent of pool sizes. The

consequence is that in such cases resource appraisers cannot always use PSRK as currently constructed. This scenario has in fact been addressed in work on translation of geologists' a priori judgments about relative likelihoods of elements of a set of geologic analogies into a probability model for discovery process modeling (Schuenemeyer and Kaufman 2005). The principal idea is to replace a single pool size distribution with a probability mixture of distinct pool size distributions. The spread of the predictive distribution of an undiscovered pool size increases, in some cases substantially.

3.2 A Combinatorial Problem

The PETRIMES pool-size-by-rank procedure takes into account the discovery record in the following way. Given MLEs of super-population parameters and the total number N of pools in the play, discovered pool sizes are ordered from largest to smallest and N-n undiscovered pool sizes are assigned to order statistics intervals or gaps. Said slightly differently, undiscovered pool size order statistics are calculated conditioned on knowledge of where in the sequence of discovered pool size order statistics undiscovered pool sizes they fit. Lee and Wang's Theorem 1 is the basis for calculation of undiscovered order statistics distributions (1985). The theorem assumes iid sampling from a super-population distribution with parameters fixed at the discovery process model, an assumption equivalent to assuming that the discoverability parameter is zero, even when discovery record based PETRIMES MLE calculations yield a positive value for it. Zhouheng Chen showed by example that explicit incorporation of discoverability into calculation of projections of undiscovered pool sizes leads to projections very different from those generated by PETRIMES (personal communication).

Given a discovery record composed of *n* pool sizes in order of discovery and a number $N > n$ of pools, pool-size-by-rank requires that *N*-*n* undiscovered pool sizes be assigned to order statistics intervals. This poses a combinatorial problem. By ordering discovery sizes from largest to smallest, they can be used to define *n*+1 discovery record intervals. Given discovery sizes $\{x_1,...x_n\}$ define $x_{(k)}$ to be the kth largest. Use the $x_{(k)}s$ to partition $(0,\infty)$ into n+1 (half open) intervals $\mathbf{(}0\mathbf{,}\, x_{(n)}\mathbf{]},...,\mathbf{(}x_{(k+1)}\mathbf{,}\, x_{(k)}\mathbf{]},...,\mathbf{(}x_{(1)}\mathbf{,}\,\infty\mathbf{)}$. A pool-size-by-rank requires the assessor to assign each of *N*-*n* undiscovered pool sizes to one of these intervals. How should this be done? How it is usually done? Each undiscovered pool size can, in principle, be assigned to any one of *n*+1 intervals so that the number of possible assignments is $(n+1)^{N-n}$. As *N*-*n* increases with either *n* fixed or in the asymptotic regime $\frac{n}{N} \equiv f_n \to f \in (0,1)$ as $N \to \infty$ the number of possible assignments becomes exponentially large. Of course, some assignments may be judged to have vanishingly small probability of occurring. In practice, a geologist often chooses exactly one assignment and then uses PETRIMES or GRASP to compute the conditional distribution of each undiscovered pool size. Once an undiscovered pool size is assigned to an order statistic interval, its probability distribution is then restricted to lie in that interval.

Let $q\mathbf{(}k\mathbf{)}$ be the number of undiscovered pools assigned to the interval $\mathbf{(}x_{\mathbf{(}k+1\mathbf{)}}\mathbf{,}\, x_{\mathbf{(}k\mathbf{)}}\mathbf{]}$ with $x_{(n+1)} = 0$ .Then the probability distribution for each of these $q\mathbf{(}k\mathbf{)}$ undiscovered pool sizes possesses a (posterior to the discovery record) domain of support restricted to $\mathbf{(}x_{(k+1)}\mathbf{,}\, x_{(k)}\mathbf{]}$. This restriction is an artifact of assignment of undiscovered pool sizes to that interval with probability one! More realistically, one should allow for assignments of undiscovered pool sizes other than $q(k)$ to each interval. According to Desselles, this is where conflicts in assessor's judgments

appear. Assignment of *N-n* undiscovered pool sizes to order statistics intervals assumes an unachievable level of certainty regarding when these pools are expected to be discovered. He notes that the timing and sizes of discoveries on the Outer Continental Shelf often break out of pool-size-by-rank restricted order statistics intervals. Technology more than geology controls these phenomena. Incremental technological advances (ultra-deep water, high pressure high temperature drilling etc.) led to large discoveries in mature plays. Desselles goes on to say that unless the *N-n* undiscovered pool sizes are properly incorporated into the discovery process model, the effect is to bias the discoverability efficiency parameter. This technology based effect needs to be integrated into the model in order to assure a reasonable estimate of the discoverability parameter.

BOEM personnel recognize that these features of pool-size-by-rank raises difficulties. The North Sea is a prime non-US offshore example. When exploration accelerated back in the 1970s and 1980s, the timing of release of lease blocks severely restricted targets available for drilling. Inferences about North Sea discoverability parameters based on straightforward sampling proportional to size and without replacement (SWORP) leads to a biased MLE of discoverability and so to bias in the projection of what remained to be discovered. Restrictions of this type can be incorporated into GRASP. The model becomes more complex. However, the real world must dictate the model—the reverse doesn't work well. North Sea computational examples show how such restrictions impact inference about the discoverability parameter (Adelman et al. 1983).

Because, in GRASP type models, undiscovered pool sizes are positively correlated. It is difficult for geologists to make coherent subjective projections of undiscovered pool sizes given a post-discovery record. In the absence of information beyond a geologist's a priori judgments about

pool sizes and prospects prior to the first discovery, the joint predictive distribution of undiscovered pool sizes summarizes all relevant probabilistic information.

3.3 An Example

To demonstrate how prediction conditioned on a discovery record using Eq. (1.5) yields results quite different from predictive quantities produced by a pool-size-by-rank, suppose that the largest pool size in the discovery record is x^* . Given discovery record $\mathbf{s}_n = (x_1 \in dx_1,.., x_n \in dx_n)$

$$Prob\{\max\{X_{n+1},.., X_N\} \geq x^* \,|\, \mathbf{s}_n\} = 1 - Prob\{X_{n+1} \leq x^*,.., X_N \leq x^* \,|\, \mathbf{s}_n\} \,. \tag{3.1}$$

For fixed values of θ, α , N and λ Eq. (1.5) and Eq. (1.8) lead to

$$Prob\{X_{n+1} \leq x^*,..., X_N \leq x^* \,|\, \lambda\} \sim F_h^{N-n}(x^* \,|\, \lambda; \theta, \alpha) \tag{3.2}$$

so unconditional as regards λ

$$Prob\{X_{n+1} \leq x^*,..., X_N \leq x^* \,|\, \mathbf{s}_n\} = \frac{1}{C(\mathbf{s}_n, \theta, \alpha)} \int_0^\infty F_h^{N-n}(x^* \,|\, \lambda; \theta, \alpha)\, dF_{Z_n}(\lambda) \tag{3.3}$$

Canadian Arctic Archipelago Western Sverdrup Basin Heiberg gas play data from Chen and Osadetz shows how sensitive largest pool sized predicative distributions are to variations in the discoverability parameter (2006). (Recall that pool-size-by-rank assigbns value zero to it when computing order statistics intervals). As of 2005, there were twenty gas discoveries and a projection of in place gas in an additional thirty-six prospects (Chen and Osadetz 2006). Treating prospect projections and discoveries as a complete finite population of N = 56 accumulation sizes

generated by a Lognormal sampling process, MLEs of Lognormal parameters are $\hat{\mu} = 1.863$, $\hat{\sigma}^2 = 1.617$. Figure 4 is a lognormal accumulation size distribution with parameters fixed at MLEs in units of cubic meters of gas. Discovered accumulation sizes are shown as tick marks on the x-axis. Gaps between tick marks can be interpreted as PSRK Module order statistics gaps. For this distribution, the probability of observing an undiscovered pool size greater than the size ($102 \text{ x} 10^6 m^3$) of the largest discovered accumulation is less than 0.012. (Fig. 4).

(Figure 4 Here)

Discovery magnitudes in order of observation are generated by successive sampling. The MLE of the discoverability parameter is $\hat{\alpha} = 0.8371$. Figure 5 displays distributions of the largest gas pool remaining to be discovered in the Heiberg play as a function of the discoverability parameter and the first twenty discoveries alone. The number *N-n* of pools remaining to be discovered is chosen to be thirty-six. Two predictive distributions for the largest pool remaining to be discovered using fixed MLEs of Lognormal pool size parameters and a variant of the PSRK module in GRASP are displayed in Fig. 5.

(Figure 5 Here)

These two distributions tell the story for α = 0 (green) the distribution of the largest pool size has observable positive probability on a large interval from 0 to $200 \times 10^6 M^3$ with a modal value at $44 \times 10^6 M^3$. For $\hat{\alpha}$ =0.8371 (black) the distribution of the largest pool size has positive probability on a much smaller interval—from 0 to approximately $50 \times 10^6 M^3$ with a mode at 8 $\times 10^6 M^3$. The red and blue lines corresponding to α = 0.3 and α = 0.6, respectively, fill in the gaps.

The BOEM adopted recommendations of an American Association of Petroleum Geologists (AAPG) committee 2007 GRASP review and removed the Pool Size Constrained by Discovery module from GRASP. The model's likelihood function should be tailored to include information about prospects and revisions of discovered pool sizes accruing after a suite of discoveries in a play. In particular, the revised likelihood function should account for new technological developments that expand the set of economically viable prospects and for lease restrictions that block prospect drilling.

An important issue is how best to incorporate information about targets for drilling not on the table at the outset of exploration but which appear as exploration progresses. Information about targets not currently drillable influences a geologist's judgments about sizes of undiscovered pools and so must influence how a geologist allocates undiscovered pool sizes to order statistics gaps. This data type requires a change in the sample frame used for inferences about discovery process model parameters.

4 Subjective GRASP, Pool-Size-by-Rank, Discovery Decline Curve and Creaming

When a discovery record assessments of pool sizes and pool counts is absent, judgments about undiscovered pool sizes are necessarily subjective. Subjective GRASP encodes a priori probability judgments about play, prospect and exploration risk, pool sizes, prospect and pool counts in a play and generates a large number of probabilistic projections. The PSRK module produces marginal probability distributions of rank order statistics generated by a random number of random pool sizes. In spite of the logical disconnect between the discovery process model and GRASP's pool-size-by-rank calculations, pool size-by-rank mean values possess a

natural discovery process model analogue: the classical discovery decline curve composed of mean values of discovery sizes in order of discovery.

Given N and θ in the special case when $L(\lambda)$ is the LaPlace transform (1.6) with $\alpha = 1$ and L' and L'' are first and second derivatives with respect to λ the expected size of the n^{th} discovery is

$$E(Y_n) = n\binom{N}{n}\int_0^\infty \frac{L''(\lambda)}{L'(\lambda)} \times [1 - L(\lambda)]^{n-1}[L(\lambda)]^{N-n}\, dL(\lambda). \tag{4.1a}$$

More generally, for $q = 1,2,...$ and $\alpha = 1$

$$E(Y_n^q) = n\binom{N}{n}\int_0^\infty \frac{L^{(q)}(\lambda)}{L'(\lambda)} \times [1 - L(\lambda)]^{n-1}[L(\lambda)]^{N-n}\, dL(\lambda) \tag{4.1b}$$

For some super-population densities equations 4.1a and 4.1b lead to exact expressions for moments. For example, if $f(x|\delta, r)$ is a Gamma density indexed by parameters r and δ then

$$E(Y_n) = \frac{r(r+1)}{\delta} \frac{\Gamma(N+1)\Gamma(N-n+\frac{1}{r}+1)}{\Gamma(N-n+1)\Gamma(N+\frac{1}{r}+1)} \tag{4.2}$$

When f is exponential with mean $1/\delta$ the decline in discovery size as a function of discovery number is linear

$$E(Y_n) = \frac{2}{\delta} \times (1 - \frac{n}{N+1}) \tag{4.3}$$

In addition

$$Cov(Y_n, Y_m) = \frac{4m(N-n+1)}{\delta^2(N+1)(N+2)} \tag{4.3b}$$

and

$$Var(Y_n) = \frac{1}{2}E^2(Y_n) + \frac{6n}{\delta^2(N+1)^2(N+2)}. \qquad (4.3c)$$

Figure 6 is a family of discovery decline curves $E(Y_n)$ versus *n* for lognormal *f* (Barouch and Kaufman, 1976).

(Figure 6 Here)

A plot of $E(S_n) = \sum_{k=1}^{n} E(Y_k)$ versus *n* is a version of a "creaming curve" popularized by Meissner and Demirmen (1981).

5 Broadening the Discovery Process Model

As stated at the outset, a 2007 review of GRASP II by an AAPG committee concluded that the pool-size-by-rank procedure understated play potential uncertainty. However, they did not identify where in the chain of GRASP computational logic this takes place. A starting point is to recognize that traditional discovery process models are limited band-width models fit to sample data consisting of pool sizes in order of discovery. How might discovery process models be broadened to incorporate richer types of geological and engineering information?

A first step is to consider how information beyond the discovery record—auxiliary information—can be used to compute a predictive distribution for undiscovered pool sizes. The model's likelihood function should be expanded to incorporate this information and bring it to bear directly on inference about the discoverability parameter, pool size distribution parameters and the number *N* of pools in the play. The pool-size-by-rank procedure currently employed in GRASP does not use auxiliary information to compute MLEs of model parameters. GRASP treats auxiliary information as if it is ancillary. (The sampling distribution of an ancillary statistic does not depend on which of the probability distributions among those being considered is the distribution of the statistical population from which the data were taken.). PETRIMES and GRASP are designed to operate on a play's discovery record consisting of discovered pool sizes

in order of discovery supplemented with assignment of a probability distribution to the total number of pools in the play. A GRASP MLE is not a function of geologists' post-discovery record judgments and, in particular prospect information is not in the sample frame.

Suppose, as considered in Sect. 3, that twenty discoveries have been made in the Heiberg play by 2005. Now consider two (imaginary) extreme scenarios. At one extreme, geologists have detailed reconnaissance information that identifies all possible prospects that might possibly be drilled in this play; in particular, the in place BOE potential of each mapped prospect is known with near certainty. At the other extreme, the only information available to geologists consists of sizes of pools discovered as of 2005. Examine the latter case first. Absent any auxiliary information at all, a geologist's judgment about play potential is shaped by the discovery record and little else. The discovery process model is then a vehicle for inferring properties of the joint predictive distribution of undiscovered pools in the absence of auxiliary information. However, once a knowledgeable geologist examines the discovery record, he forms opinions about remaining play potential. These opinions are not pre-discovery record opinions and so should not be processed via Bayes' Theorem by assigning a judgmental prior distribution to model parameters prior to observing any drilling outcome in the play at all → observe data → compute posterior distribution of parameters → compute predictive distribution of play potential. The geologist's judgments are confounded with (some would say contaminated by) his observation of and interpretation of the discovery record. The analyst who wishes to adhere to probability logic faces a few difficulties. Once a geologist's judgments about play potential are confounded by observation of the data (the discovery record) how should his judgments and the discovery record be combined? What sort of predictive distribution should be computed? The pool-size-by-rank procedure cuts this Gordian knot but, unfortunately, in a fashion that is not always

probabilistically coherent. Turn next to the case when, in addition to the discovery record, the number of prospects along with each prospect's BOE potential is known with certainty. The only major uncertainty remaining is the order in which prospects may be drilled and which prospects are in fact pools. The need for a super-population process pool size distribution disappears. Successive sampling applied to a finite population of prospect sizes and discovered pool sizes, each treated as known with certainty, is then a relevant model to project discovery order of undiscovered pools.

As exploration of a play unfolds, an increasingly rich set of geologic and engineering data accumulates. The play's boundary may expand or contract and sizes and locations of new drillable prospects may be identified. Auxiliary information of this sort clearly influences how a geologist chooses an allocation of undiscovered pool sizes to order statistics gaps in the discovery record. But it does so in a completely informal way and is, to our knowledge, not used to revise judgments about model parameters before implementation of the pool-size-by-rank procedure. Another team of assessment geologists could make a different assignment of pool sizes into gaps. Revision of judgments about model parameters should be based on all sample information available. In their study of Canada's Western Sverdrup Basin play Chen and Osadetz (2006) present prospect data that, along with the discovery record, can be used to this end. The model should be reformulated so that prior to any exploratory drilling at all the model embraces both tested and abandoned prospects pools and accumulations. In addition, they introduce spatial modeling of prospects and discoveries. Rabinowitz (1991) studied a version of this generalized successive sampling problem. Introduction of uncertainty about parameters, such as the number of pools and prospects in a play, as done in the PSRK module is not equivalent to expanding the likelihood function to embrace prospect data and spatial characteristics of deposition.

An enriched sample frame that includes tested and abandoned prospects along with discovery sizes in order of observation is valuable. Chen and Osadetz (2006) recommend that dry holes be replaced with tested and abandoned prospects, as a more accurate description. At the outset, the set of targets for drilling consists of two objects: a set of dry prospect sizes and a set of prospect sizes, each of which when drilled turns out to be a productive pool (field) or a tested and abandoned prospect. It is reasonable to assume that both prospect sizes projected by geologists and productive pool sizes are generated by a super-population process. An essential distinction among GRASP, PETRIMES and a model designed to capture auxiliary information is that the last of these adopts sampling proportional to size and without replacement from a finite population composed of the union of a realization of a set of pools size and a set of prospect sizes that when tested will be abandoned.

Modifying the likelihood functions used in GRASP and PETRIMES type models to account for temporal effects on exploratory drilling imposed by lease blocking and technology breaks would substantially increase predictive validity. Considerable recent progress has been made on spatial modeling of features of oil and gas fields and prospects that influence discoverability. For example, Chen and Osadetz (2006) define the set of targets for drilling to be the union of a set of dry prospect sizes and a set of prospect sizes each of which when drilled turns out to be a productive pool (field) or a tested and abandoned prospect. Elements of the union of these can two sets are treated as nodes in a spatial network with arcs that encode geologic and/or dependencies among them. Martinelli et al. (2011) show how Bayesian networks can be deployed to this end. Discovery process models that combine spatial networking and basic principles of discovery process modeling are an important next step.

Acknowledgments The authors wish to thank Richard Desselles for constructive criticism and Zhouheng Chen, Kirk Osadetz and Walt Stromquist for valuable commentary as well as Jennifer Challis for her invaluable editorial assistance.

References

Adelman MA, Houghton J, Kaufman GM, Zimmerman M (1983) Energy resources in an uncertain future: coal, gas, oil, and uranium supply forecasting. Ballinger Publishing Co, Cambridge, 248-261

Andreatta G, Kaufman GM (1986) Estimation of finite population properties when sampling is without replacement and proportional to magnitude. Journal of the American Statistical Association, 81:657-666

Barouch E, Kaufman GM (1976) Oil and gas discovery modeled as sampling proportional to size. Sloan School of Management Working Paper, No. 888-76, 1-64

Chen Z, Osadetz K (2006) Geological risk mapping and prospect evaluation using multivariate and Bayesian statistical methods, western Sverdrup Basin of Canada. AAPG Bulletin, 90:859–872

Kaufman GM, Balcer Y, Kruyt D (1975) A probabilistic model of oil and gas discovery. Studies in Geology No. 1 - Methods of Estimating the Volume of Undiscovered Oil and Gas Resources. The American Association of Petroleum Geologists Haun, J.D. Ed.

Kaufman GM, Chow C, Park C (1988) Estimation of a discoverability parameter when sampling a finite population successively. Brookhaven Contract Report

Lee PJ (2008) Statistical methods for estimating petroleum resources. Oxford University Press, New York, 256 p

Lee PJ, Wang PCC (1983a) Probabilistic formulation of a method for the evaluation of petroleum resources. Math Geol 15:163-181

Lee PJ, Wang PCC (1983b) Conditional analysis for petroleum resource evaluations. Math Geol 15:349-361

Lee PJ, Wang PCC (1985) Prediction of oil or gas pool sizes when discovery record is available. Math Geol 17:95-113

Lee PJ, Wang PCC (1986) Evaluation of petroleum resources from pool size distribution. In: Rice DD, Oil and gas assessment – Methods and applications. Am Assoc Petroleum Geologist Studies in Geology, Tulsa, OK, AAGP, 21:33-42

Lee PJ, Tzeng HP (1993) The petroleum exploration and resource evaluation system (Petrimes): Working reference guide Version 3.0 (PC version). Institute of Sedimentary and Petroleum Geology, Geological Survey of Canada Open File Report 2703. Calgary, Canada, p 204

Lore GL, Ross KM, Bascle BJ, Nixon LD, Klazynski RJ (1991) Assessment of conventionally recoverable oil and gas resources of the Gulf of Mexico and the Atlantic Outer Continental Shelf. OCS Report MMS 99-0034 U.S. Minerals Management Service, New Orleans, LA

Martinelli G, Eidsvik J, Hauge R, Drange-Forland M (2011) Bayesian networks for prospect analysis in the north sea. AAPG Bulletin. 95:1423–1442

McCrossan, R.G., R.M.Procter and W.J. Ward(1981). "Estimate of oil resources, Lloydminster Area, Alberta," Proceedings of the First International Conference on the Future of Heavy Crude and Tar Sands"

Meisner, J. and Demirmen,F. (1981) The creaming method: a Bayesian procedure to forecast future oil and gas discoveries in mature exploration provinces: Journal of the Royal Statistical Society, v. 144, part A, p. 1-31.

Miller, BM, Thomsen HL, Dolton GL, Coury AB, Hendricks TA, Lennartz FE, Powers RB, Sable EG, Varnes KL (1975) Geological estimates of undiscovered recoverable oil and gas resources in the United States. U.S. Geol. Survey Circ. 726

Rabinowitz D (1991) Using exploration history to estimate undiscovered resources. Math Geol 23:257-274

Schuenemeyer, JH, Kaufman GM (2004) The New Frontier – A Probabilistic Assessment of the Petroleum Potential of the Circum Arctic, Ilulissat, Greenland.

Stromquist W (1998) Review of the Discovery Process Approach in the PETRIMES resource estimation software and Review of the Subjective Approach in PETRIMES. Report to MMS

West M (1994) Discovery Sampling and Selection Models. In: JO Berger and SS Gupta (eds) Decision Theory and Related Topics IV. Springer, Verlag New York, 221-235

West M (1996) Inference in successive sampling discovery models. Journal of Econometrics. 75:217-238

Appendix A: Undiscovered Pool Sizes are Positively Correlated

Assertion: Undiscovered pool sizes conditional on a discovery record are positively correlated.

Proof: Define $E(X_k|\lambda;\theta,\alpha)$ to be the expectation of a generic $X_k, k=n+1,...,N$ conditional on $\Lambda=\lambda$, pool size parameters θ and discoverability parameter α. For $n+1\le k,m\le N$ the covariance $Cov(X_k,X_m|\theta,\alpha)$ of X_k and X_m conditional on pool size parameters θ and discoverability parameter α is, using the covariance decomposition formula

$$Cov(X_k,X_m|\theta,\alpha)=E_\lambda Cov(X_k,X_m|\lambda;\theta,\alpha)+Cov_\lambda(E_\lambda(X_k),E_\lambda(X_m)) \quad (A.1)$$

As X_k and X_m are conditionally independent given $\Lambda=\lambda$,

$$Cov(X_k,X_m|\theta,\alpha)=Cov_\lambda(E_\lambda(X_k),E_\lambda(X_m)). \quad (A.2)$$

Here Λ is a rv with range set $(0,\infty)$. Chebychev's well known functional inequality says that if g and h are functions with common domain, then, if both g and h are strictly increasing or both are strictly decreasing as λ traverses $(0,\infty)$ then

$$\int_0^\infty g(\lambda)h(\lambda)dF_\Lambda(\lambda)>\int_0^\infty g(\lambda)dF_\Lambda(\lambda)\times\int_0^\infty h(\lambda)dF_\Lambda(\lambda). \quad (A.3)$$

Both $E(X_k|\lambda;\theta,\alpha)=E_\Lambda(X_k)$ and $E(X_m|\lambda;\theta,\alpha)=E_\Lambda(X_m)$ are strictly decreasing as λ traverses $(0,\infty)$. Hence

$$\int_0^\infty E_\Lambda(X_k)\times E_\Lambda(X_m)dF_\Lambda(\lambda)>\int_0^\infty E_\Lambda(X_k)dF_\Lambda(\lambda)\times\int_0^\infty E_\Lambda(X_m)dF_\Lambda(\lambda) \quad (A.4)$$

so $Cov(X_k,X_m|\theta,\alpha)$ ■

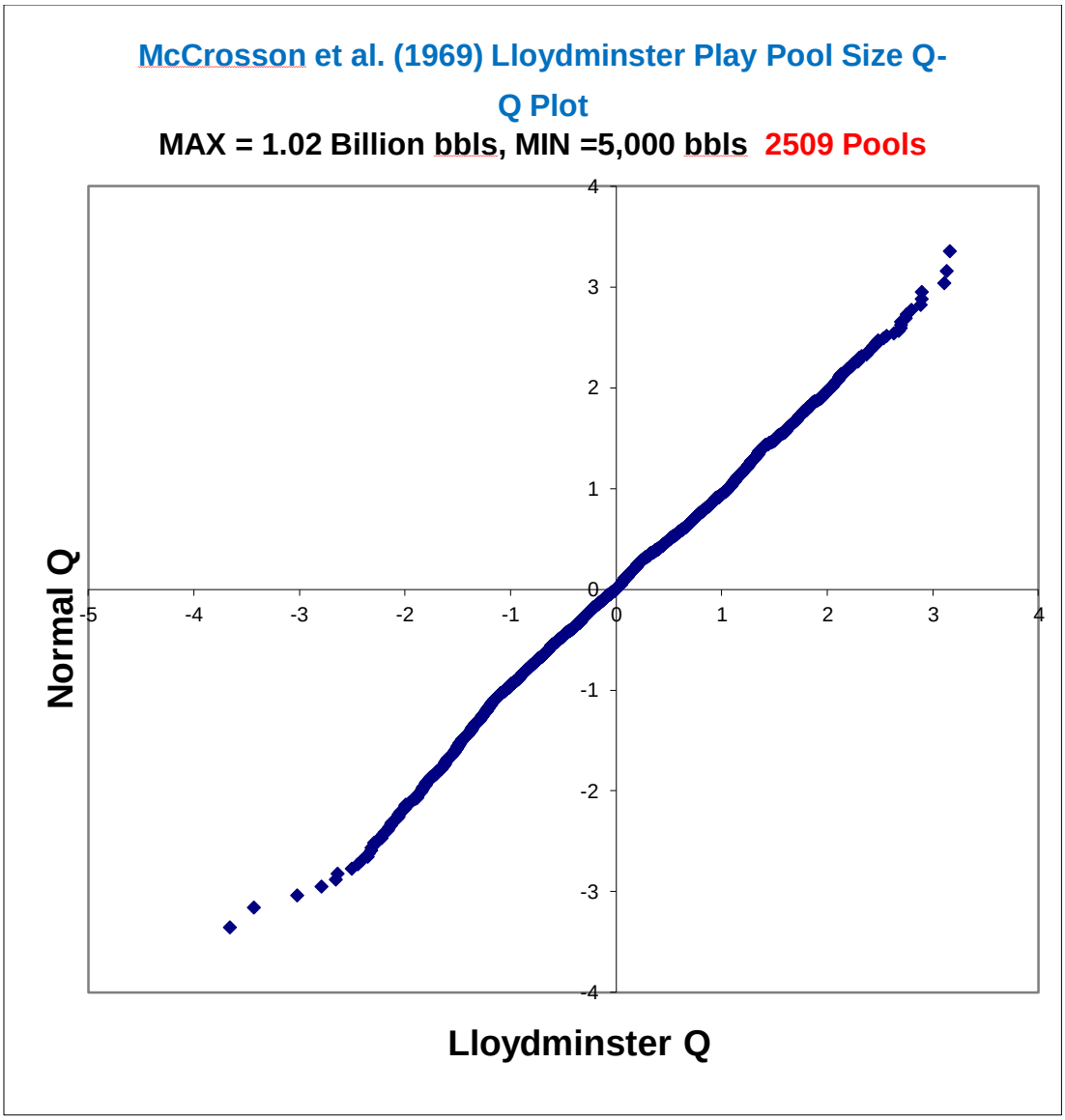

Figure 1. Q-Q plot of log sizes versus a Normal distribution

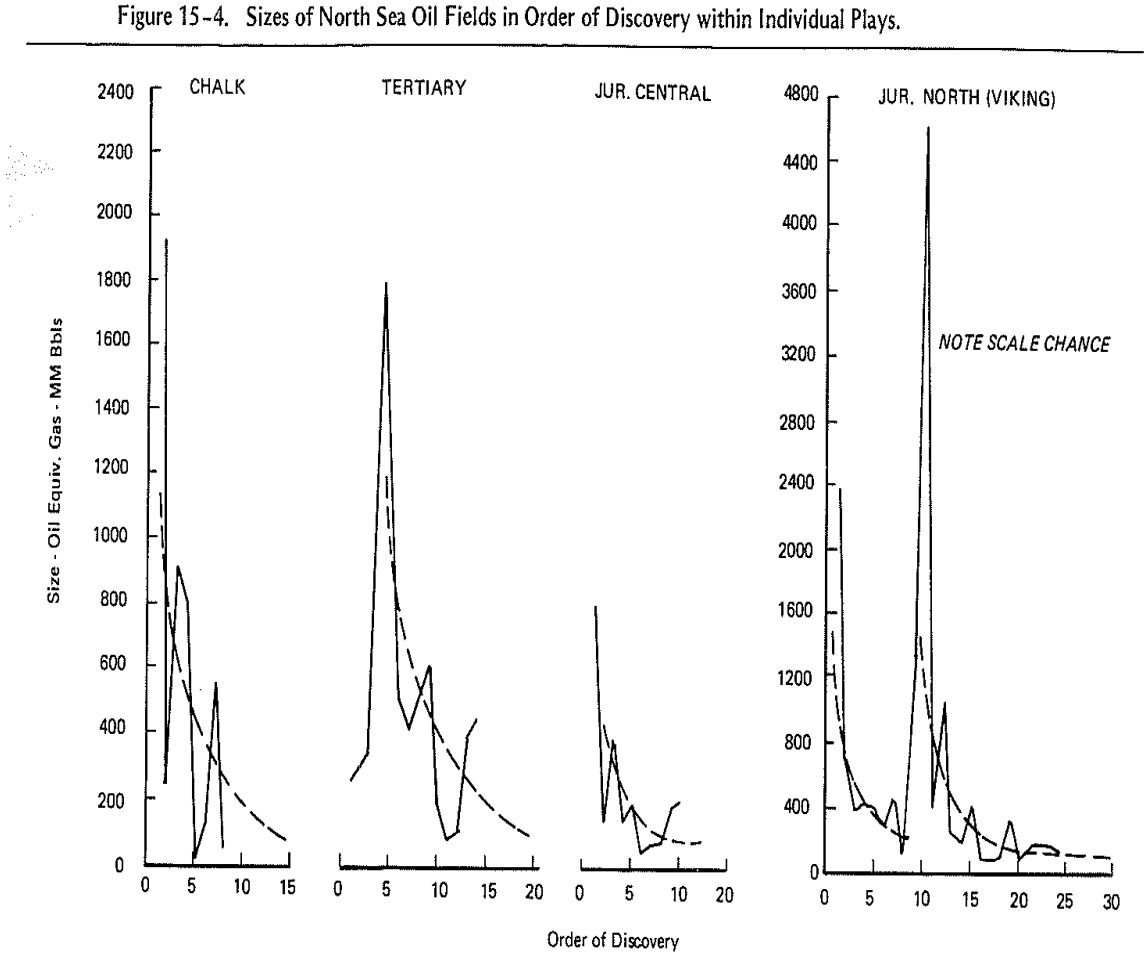

Figure 2. North Sea oil field sizes in order of discovery within individual plays (from Adelman et al., 1983)

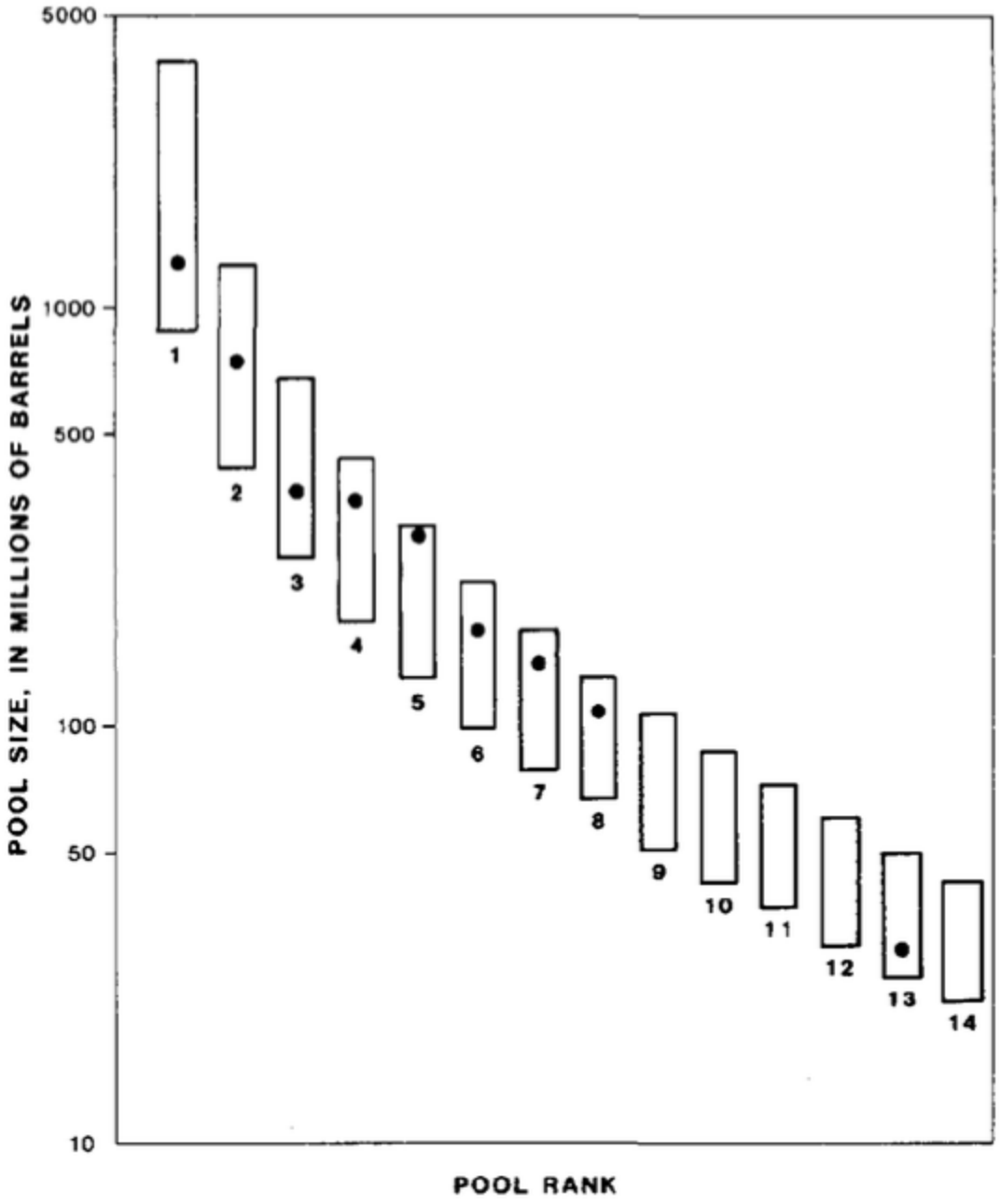

Figure 3. Predicted pool sizes by rank for the Rimbey-Meadowbrook reef play. Dots indicate reserves of pools and boxes indicate values at the 25th and 75 percentiles (Lee and Wang, 1985, Fig. 8)

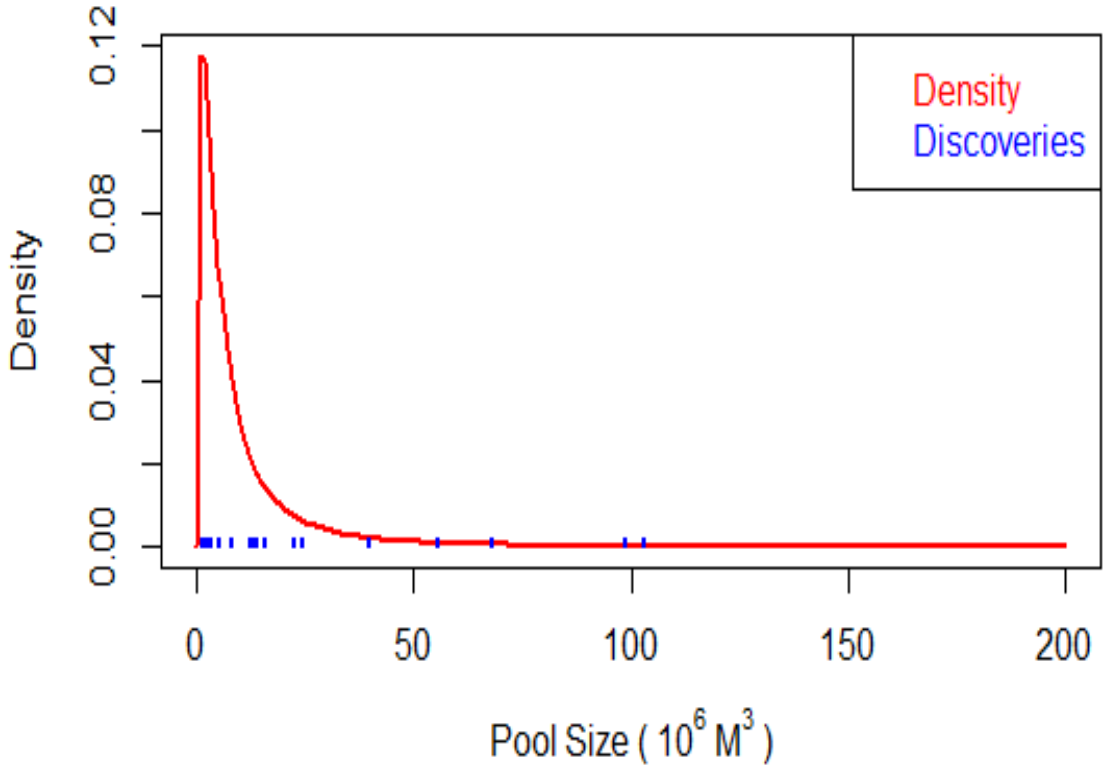

Figure 4. Lognormal density for the Sverdrup play (red) and discovered pools (blue)

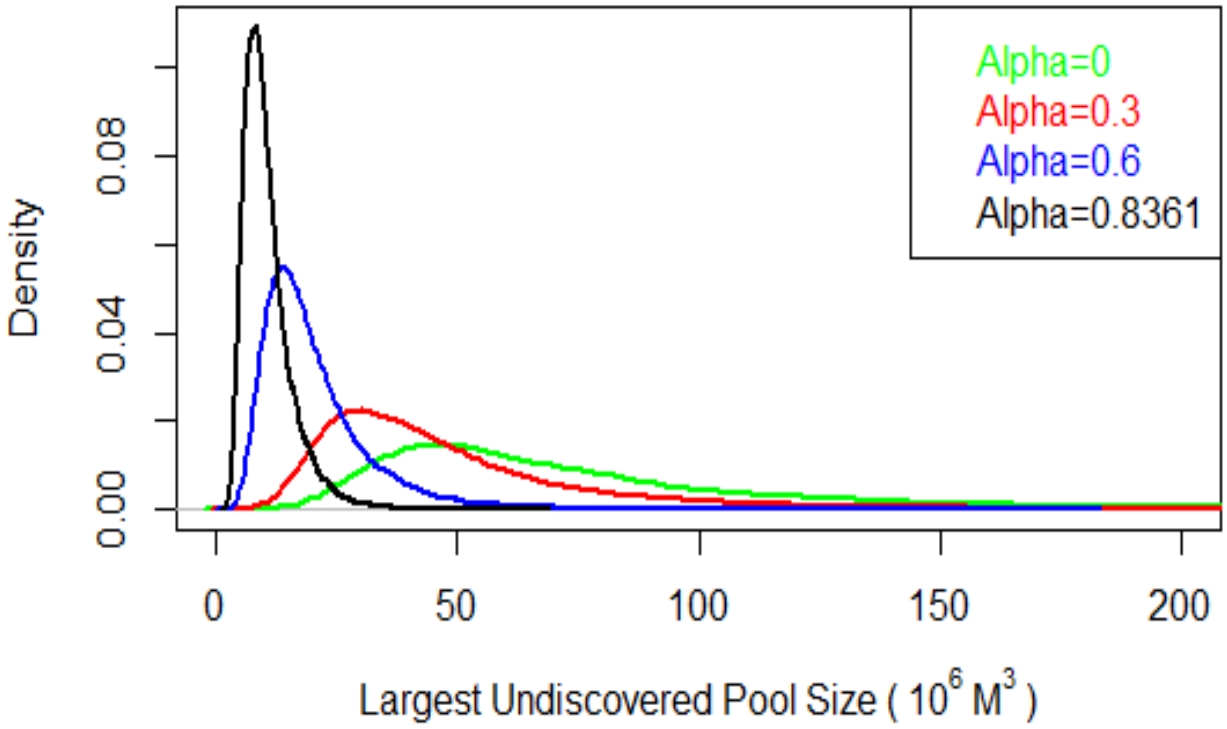

Figure 5. Distributions of the largest undiscovered pool size for four values of the discoverability parameter

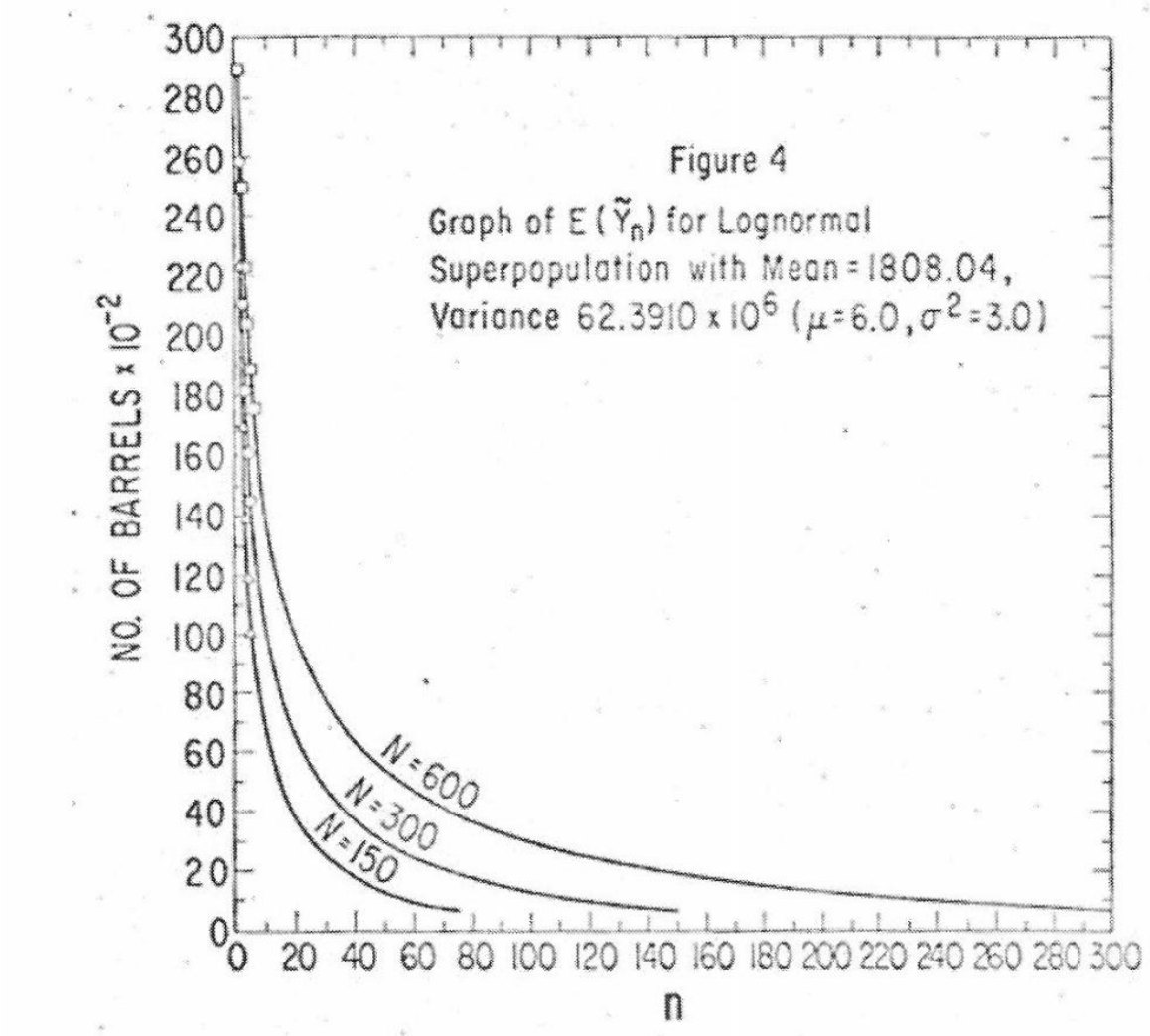

Figure 6. Family of decline curves for a Lognormal Super-population