NTIS: HC # 3.00

TWO STOCHASTIC MODELS USEFUL IN PETROLEUM EXPLORATION

by

Gordon M. Kaufman
Sloan School of Management
Massachusetts Institute of Technology

and

Paul G. Bradley
Department of Economics
University of British Columbia

written for the

Second International Symposium on Arctic Geology

This work was supported in part by a grant from Resources for the Future. A portion of G.M. Kaufman's support came from NASA Contract No. NGL-22-009-309, Integrated Planning and Control Systems. P.G. Bradley was the holder of a Canada Council Postdoctoral Fellowship.

19

Introduction

The process of exploring for oil and gas is rich in uncertainties. Any attempt to forecast returns to investment in exploration must take them into account in a systematic way. By this we mean that inferences about the important uncertain quantities characterizing the exploration process should be based on a mathematical model whose parameters may be estimated from observable data in a coherent way. At the root of any useful model of the exploration process, then, is a set of assumptions that delineate in clear unambiguous terms the probability law governing the manner in which observable data is generated.

Our first objective is to construct a model of the exploration process that allows us to test empirically the hypothesis that at an early stage in the exploration of a basin, the process behaves like sampling without replacement. The model we posit is parsimonious -- based on a small number of assumptions and indexed by only five parameters. The set of assumptions on which it is built reflects at least two qualitative assertions often made by oilmen: the "big ones" tend to be found first and the size distribution of fields is highly skewed. We may use it to compute answers to two questions of paramount importance in designing exploration strategy:

(1) How does the probability that a wildcat well will find a reservoir change (if at all!) as the history of a basin unfolds?

(2) What is the probability that a yet-to-be-drilled wildcat well will find a reservoir of a given size or greater at a given point in the development of a basin?

Our second objective is to posit a reasonable model of the spatial distribution of petroleum reservoirs that conforms to a number of empirically observed facts about such distributions, but does not possess three unrealistic attributes that characterize models of spatial occurrence appearing in the literature: dependence of the model on arbitrary subdivision of a basin into units of subspace, the assumption of spatial homogeneity of the stochastic process operating within each such unit as well as across units, and conceptualization of a reservoir as a point (in the plane) rather than as an object with positive area. (See Uhler and Bradley [1970], Allais [1957], Engel [1957].)

The first model we pose differs significantly from those postulated by Arps and Roberts [1956], and by Kaufman [1963]. It accounts for the impact of exploration technology on the probability of discovering a new reservoir in an explicit and intuitively meaningful way. And it is structured so that inferences about parameters not known with certainty may be made in accordance with well understood statistical principles. In particular, the assumption that the probability of discovering a reservoir is proportional to its size strongly biases any "usual" estimator when the sample size is small, so we develop methods for coping with this complicating feature of the data-generating process.

Our spatial model has not yet been subjected to empirical validation. However, its structure is sufficiently flexible to warrant the conjecture that it will in fact prove to be a reasonable characterization of a process that can by visual inspection be seen to be spatially inhomogeneous; i.e., fields tend to cluster rather than to be spread evenly

throughout a basin. Under the direction of one of the authors, Golovin [1970] has programmed versions of this model and done computational exploration of some of its features. We shall draw heavily on his work in our discussion.

I. A Model of the Discovery Process

Nomenclature

The technology available for identifying potential oil-and/or-gas-bearing structures is not perfect. We shall assume that if this technology is applied to the entire areal extent of a generic basin it will delineate M distinguishable prospects. We label them 1,2,...,M and call $\mathcal{M} = \{1,2,\ldots,M\}$ the label set for the population of prospects in this basin. Each prospect either is or isn't a field; by "field" we mean a hydrocarbon-bearing reservoir or a collection of contiguous reservoirs. (Precision in defining what we mean by a "field" is not important at this juncture.) If it is a field, the field has many characteristics of interest; momentarily, we focus on only one -- its areal extent.

Let

$$x_i = \begin{cases} 1 & \text{if the } i^{th} \text{ prospect is a field} \\ 0 & \text{otherwise} \end{cases}$$

and define

$$A_i = \text{areal extent of } i^{th} \text{ prospect.}$$

Then (x_i, A_i) for $i \in \mathcal{M}$ is a characteristic of the i^{th} population element. We do not know $\Theta_M \equiv \{(x_i, A_i) \mid i \in \mathcal{M}\}$ with certainty prior to beginning exploration of the basin. One of our objectives is to make inferences about Θ_M as prospects are delineated and fields discovered. In particular we wish to know which elements of Θ_M have $x = 1$, since the i^{th} prospect is by definition a field if and only if $x_i = 1$.

At the outset of exploration, the exploration process will generate only a small subset of potential prospects in the basin, say

$n < M$ of them with population labels $i_1,\ldots,i_n$. And only a subset of $k < n$ of these prospects will have been drilled. Hence a sample of size n is an ordered sequence of n of the population elements, $(i_1,\ldots,i_n)$ with $i_\ell \in \mathcal{M}$ for $\ell = 1,2,\ldots,n$, together with an ordered n-tuple of observed characteristics; e.g.,

$$[(i_1,\ldots,i_n);\ (x_{i_1},A_{i_1}),\ A_{i_2},\ A_{i_3},\ (x_{i_4},A_{i_4}),\ldots,(x_{i_n},A_{i_n})].$$

There will be no loss in generality in the context of the model we deal with here if we relabel those prospects that have been drilled in the order in which they were drilled and re-order as follows:

$$[(i_1,\ldots,i_n);\ (x^{(1)},A^{(1)}),\ldots,(x^{(k)},A^{(k)}),\ A_{i_2},\ A_{i_3},\ldots)]$$

In fact our model will allow us to ignore the ordering of areas A_{i_j} of prospects that have been generated at a given point in time but not drilled, so we define a sample as:

$$H^r_{n,k} = [(i_1,\ldots,i_n);\ (x^{(1)},A^{(1)}),\ldots,(x^{(k)},A^{(k)});\ \{A_{i_j}\}\ ]$$

where it is understood that the element $\{A_{i_j}\}$ is the set of areas of undrilled prospects generated by the exploration process at the instant when the $(k+1)^{st}$ well is to be drilled; r is $\sum_{i=1}^{k} x^{(t)}$, the number of fields found by the first k wells. We shall use $H^r_{n,k}$ as shorthand for a complete description of a sample when no ambiguity will arise.

In order to describe the assumptions on which our model is based, we need the following array of notational ammunition:[1]

---

1. A summary of symbols is given at the end of the paper in Table 1.

$I_N = \{i \mid i \in \mathcal{M} \text{ and } x_i = 1\}$, the label set of fields in the basin,

$S_N = \sum_{i \in I_N} A_i$, the total area of N fields in the basin,

$R_M = \sum_{i=1}^{M} A_i$, the total area of M prospects in the basin,

$J_k = \{t \mid x^{(t)} = 1 \text{ for } t=1,2,\ldots,k\}$, the label set of successful wells among the first k wells drilled,

$\bar{J}_k = \{t \mid x^{(t)} = 0 \text{ for } t=1,2,\ldots,k\}$, the label set of unsuccessful wells among the first k wells drilled,

$s_k = \sum_{t \in J_k} A^{(t)}$, the total area of fields discovered by the first k wells, and

$u_k = \sum_{t=1}^{k} A^{(t)}$, the total area of prospects drilled by the first k wells.

The Data-Generating Model

We shall assume that the process generating observable data has the following properties:

1. Constant Technology

Given S_N and R_M and conditional on observing a sample $H^r_{n,k}$ yielding statistics s_k and u_k,

$$P\left(x^{(k+1)} = 1 \mid H^r_{n,k}\right) = \frac{S_N - s_k}{R_M - u_k} .$$

This assumption says that the probability that the $(k+1)^{st}$ well will discover a field changes in a "hypergeometric-like" fashion with changes

in s_k and u_k. The ratio S_N/R_M does not depend on either s_k or u_k and is a rough measure of technological efficiency, hence the label "constant technology".

2. Probabilistic Proportionality

Given $\{A_i \mid i \in \eta$ and $x_i = 1\}$ and conditional on observing $H^r_{n,k}$ and $\tilde{x}^{(k+1)} = 1$, the probability that the $(k+1)^{st}$ well discovers a field of areal extent A is

$$P\left(\tilde{A}^{(k+1)} = A \mid \tilde{x}^{(k+1)} = 1,\ H^r_{n,k}\right) = \begin{cases} \dfrac{A}{S_N - s_k} & \text{if } A \in \{A_i \mid x_i = 1 \text{ and } i \notin J_k\} \\ 0 & \text{otherwise.} \end{cases}$$

Assumptions 1 and 2 formalize the idea that the probability of discovering a field of areal extent A is proportional to A, for given R_M and S_N,

$$P\left(\tilde{A}^{(k+1)} = A,\ \tilde{x}^{(k+1)} = 1 \mid H^r_{n,k}\right) = \begin{cases} \dfrac{A}{R_M - u_k} & \text{if } A \in \{A_i \mid x_i = 1 \text{ and } i \notin J_k\} \\ 0 & \text{otherwise.} \end{cases}$$

Both assumptions ignore the information content of the statistic $\{A_{i_j}\}$, the set of areas A_{i_j} of prospects generated prior to drilling the $(k+1)^{st}$ well but as yet undrilled, and exploit only the information generated by the outcome of drilling the first k wells. In order to exploit all information in $H^r_{n,k}$, we would have to build a model of the process generating prospects as well as of one generating discoveries. We have chosen to suppress this complicating feature in our preliminary investigation.

3. Probability Law of $\{\tilde{A}_i \mid i \in I_N\}$

$\{\tilde{A}_i \mid i \in I_N\}$ is a set of mutually independent identically distributed random variables, each characterized by a density $f(\cdot \mid \underline{\theta})$

Likelihood Function

The likelihood function generated by observation of a sample $H^r_{n,k}$ is, defining $u_o = 0$ and $s_o = 0$,

$$L\,(N, R_M, \underline{\theta}, S_N \mid H^r_{n,k})$$

$$\propto \prod_{t=1}^{k} \left(\frac{S_N - s_{t-1}}{R_M - u_{t-1}}\right)^{x(t)} \left(1 - \left[\frac{S_N - s_{t-1}}{R_M - u_{t-1}}\right]\right)^{1-x(t)}$$

$$\times \prod_{t \in J_k} \left[\frac{A(t)}{S_N - s_{t-1}}\right] f(A^{(t)} \mid \underline{\theta}) \qquad (1.1)$$

$$\times f^{*\,N-r}\,(S_N - s_k \mid \underline{\theta})$$

where $f^{*\,N-r}$ is the (N-r)-fold convolution of f with itself. The appearance of the term $f^{*\,N-r}\,(S_N - s_k)$ may be explained like this: the process of generating observations does so in two stages. First, nature generates N values $\{A_i \mid i \in I_N\}$. Then the observables are generated in a way that depends probabilistically on $S_N = \sum_{i \in I_N} A_i$. Consequently, S_N is a parameter of the observational process (1 and 2) and at the same time a statistic from the vantage point of the process generating field areas (3). If we wish to make inferences about N, R_M, $\underline{\theta}$, and S_N jointly, then S_N appears in both roles.

The likelihood function (1.1) may be rewritten as proportional to:

$$\prod_{t \in J_k} \left(1 - \left[\frac{S_N - s_{t-1}}{R_M - u_{t-1}}\right]\right)$$

$$\times \prod_{t \in J_k} \left(\frac{A^{(t)}}{R_M - u_{t-1}}\right) \quad f(A^{(t)} \mid \underline{\theta}) \qquad (1.2)$$

$$\times f^{*N-r}(S_N - s_k \mid \underline{\theta})$$

Approximation of Likelihood Function

In general, working directly with $L(N, R_M, \underline{\theta}, S_N \mid H^r_{n,k})$ is difficult. However, when N-r is very large we can apply the (equal components) Central Limit Theorem; i.e., if f has mean $m \in (-\infty, +\infty)$ and bounded variance $\mathcal{U}$, then as N-r increases, f^{*N-r} becomes more and more accurately approximated at each value of its domain by a Normal density $f_N(\cdot \mid m[N-r], \mathcal{U}[N-r])$ with mean $m[N-r]$ and variance $\mathcal{U}[N-r]$.[2]

Here we are interested in the behavior of L when f is a Lognormal density with parameter $\underline{\theta} = (\mu, \sigma^2)$:

$$f(x \mid \underline{\theta}) = f_L(x \mid \mu, \sigma^2) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\log_e x - \mu)^2/\sigma^2} \frac{1}{x} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \qquad (1.3)$$

Combining the Normal approximation suggested above with f as in (1.3), that portion of (1.1) involving μ and σ^2 may be written as proportional to

---

2. Provided $\int |f(\xi \mid \underline{\theta})|^2 d\xi < \infty$.

$$\sigma - r_e - \tfrac{1}{2}k(g-\mu)^2/\sigma^2 - \tfrac{1}{2}\, v/\sigma^2$$

$$x(U[N-r])^{-\frac{1}{2}} e^{-\frac{1}{2}([S_N - s_k] - m[N-r])^2/U(N-r)} \qquad (1.4)$$

where

$$m = \exp\{\mu + \tfrac{1}{2}\sigma^2\}, \qquad U = m^2[\exp\{\sigma^2\} - 1],$$

$$g = \frac{1}{r}\sum \log A_i, \qquad v = \sum (\log A_i)^2 - rg^2.$$

Maximum Likelihood Estimation

It will be convenient to work with m in place of μ in the sequel and we shall do so. To find a maximum likelihood estimator (MLE) of parameters m, σ^2, N, R_M, and S_N when the likelihood function is of the form (1.1) is analytically difficult. We employ the following procedure:[3]

1. Fix the value of R_M.
2. Find an MLE $m^*(\sigma, N, S_N)$ of m for fixed σ, N, and S_N.
3. Holding N and S_N fixed, substitute $m^*(\sigma, N, S_N)$ for m in (1.4); find MLE's of m and σ^2 by searching (1.4) over $\sigma^2 \in (0, \infty)$. Call this pair $[m_*(N, S_N), \sigma^2_*(N, S_N)]$.

---

3. In practice we have utilized the gradient method developed by Goldfeld, et al.[1966] to simultaneously estimate μ (or m) and σ conditional upon the pair (N, S_N). This creates the tableau described in step 4. It may prove possible to employ this method to simultaneously estimate all parameter values, thus eliminating the search procedure of steps 4-7. Using data on exploratory drilling in Alberta, we have estimated parameters for several regions. The data support the hypothesis that the sizes of discoveries tend to decrease over time, but although the estimates appear reasonable we regard them as too tentative to be published at this time.

4. Repeat step 3 for a large set of values of the ordered pair (N,S_N), and tabulate the value of log likelihood for each (N,S_N) at $[m,\sigma^2] = [m_*(N,S_N), \sigma_*^2(N,S_N)]$.
5. Search tabulated values of the log likelihood for an approximate maximizer $(N^*,S_N^*,m_*(N^*,S_N^*),\sigma_*^2(N^*,S_N^*))$ of (1.1), given R_M.
6. Repeat steps 2 through 5 for a set of values of R_M.
7. Search log likelihood values for an approximate (joint) MLE of all parameters.

II. A Spatial Model

By a spatial model of the deposition of petroleum deposits, we mean a stochastic process generating values of a sequence of random variables in a way that jointly simulates the frequency distribution of areal extent, the geographic location and the shape of these deposits. The first approaches that pop into one's mind are incorrect; i.e., viewing the process generating the number of fields per unit area A as a spatially homogeneous Poisson process is incorrect; randomizing the parameter $\lambda(A)$ of such a process by assuming that $\lambda(A)$ is a random variable with Gamma density (see Uhler and Bradley [1970]) leads to a better approximation, but still is deficient in the tails -- that is, a negative binomial distribution doesn't fit well in the right tail. In addition, a compound Poisson process, or a (randomized) modification of it doesn't really explain the "clustering close together" that one observes when examining a map pinpointing oil and gas fields, already discovered in a well-explored basin.

The model we propose here is conceptually simple, extremely flexible and can be easily modified in many ways. We replace the two dimensional continuum with the lattice $L = \{(i,j) \mid i,j \text{ integer}\}$ of ordered pairs of integers and equip it with the simplest of probabilistic laws of motion: a symmetric random walk. We then define an imbedded process that lays down a 1 or a 0 at first (or subsequent) passage of the random walk through a lattice point. The assumptions we detail shortly lead to pictures such as that shown in Figure 1 (Golovin [1970], p. 17).

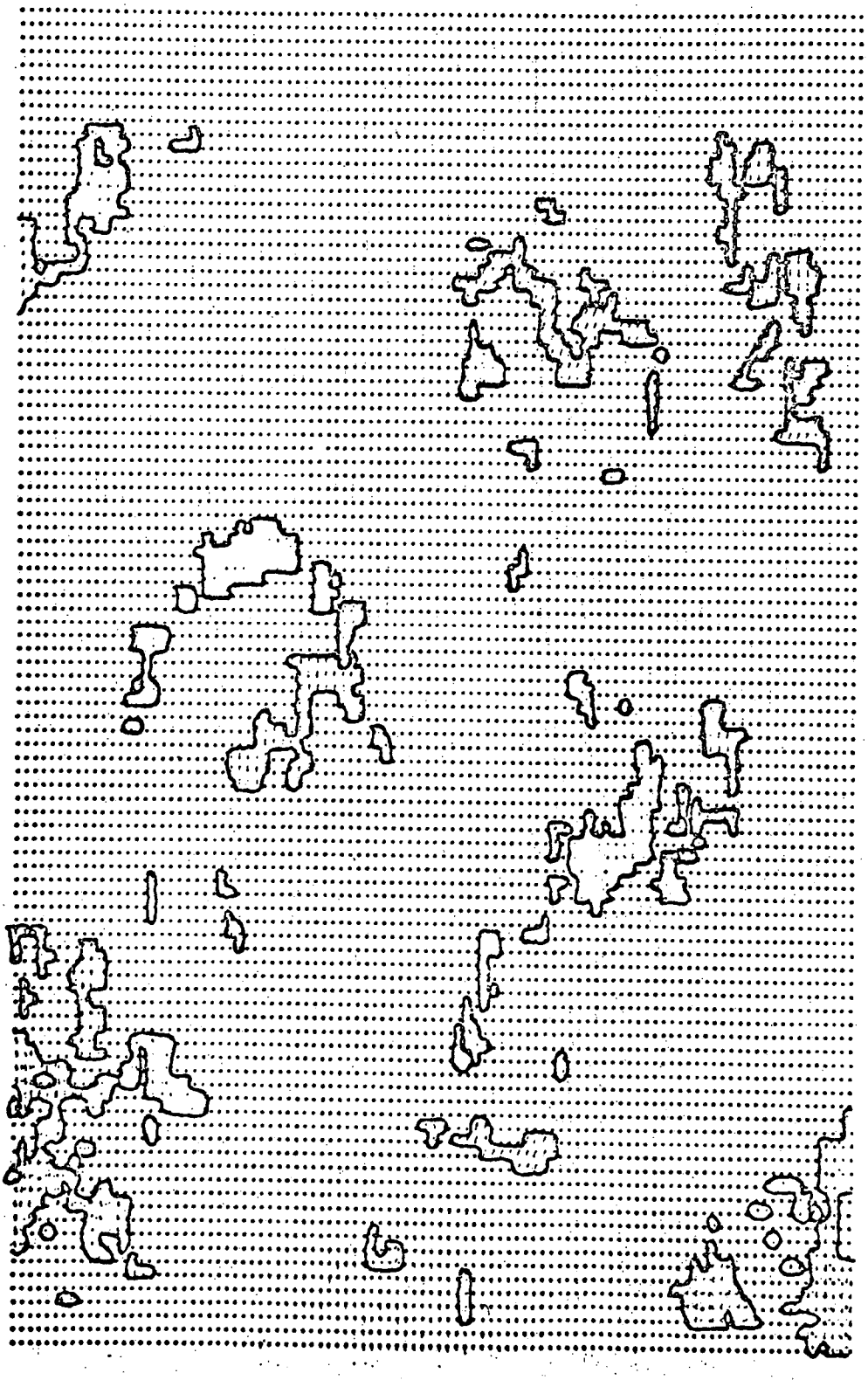

Figure 1

Distinguishing features of the model used to generate Figure 1 are that reservoirs have positive area, there is a cluster effect, and the frequency histogram of area extents is, aside from a truncation effect induced by clustering, asymptotically Lognormal.

Basic Definitions and Properties

The model is composed of three basic objects: a symmetric random walk on L, a random process superposed on the path taken by the random walk, and a stopping rule.

Let $[i,j;\ t]$ denote the position on L of the random walk at trial t, $t = 0,1,2,\ldots$ and define

$$\delta(i,j) = \begin{cases} 1 & \text{if } (i,j) \text{ has been assigned a one at some } t' \leq t, \\ 0 & \text{if } (i,j) \text{ has been assigned a zero at some } t' \leq t. \end{cases}$$

If the random walk has not passed through (i,j) at some $t' \leq t$, $\delta(i,j)$ is left undefined. We set

$$I(t) = \{(i,j) \mid \delta(i,j) = 1 \text{ at trial } t\}$$

and

$$J(t) = \{(i,j) \mid \delta(i,j) = 0 \text{ at trial } t\} ,$$

and define the state S_t of the process at trial t as a triplet consisting of the location $[i,j;\ t]$ of the random walk at the end of trial t, the set I(t), and the set J(t); i.e., $S_t = ([i,j;\ t], I(t), J(t))$. Let h_t be the smallest non-negative integer such that $[i,j;\ t+h_t] \notin I(t) \cup J(t)$; $t+h_t$ is the first trial following trial t at which first passage through an

an unassigned point occurs. Set $t_0 = 0$, $t_k = \sum_{i=1}^{k} h_i$, $k \geq 1$, and define on $t_0, t_1, \ldots, t_k, \ldots$ a sequence $\{\tilde{\xi}_{t_k}, k = 1,2,\ldots\}$ of mutually independent random variables with common probability function

$$P\{\tilde{\xi}_{t_k} = j\} = \begin{cases} 1 - \mu & \text{if } j = 0; \\ \mu\binom{m}{r}\rho^r(1-\rho)^{m-r} & \text{if } j = 2^r,\ r = 0,1,2,\ldots,m; \\ 0 & \text{otherwise;} \end{cases}$$

with m a positive integer and $0 < \rho < 1$. The value ξ_{t_k} of $\tilde{\xi}_{t_k}$ may be interpreted as a "chain" of ones that the process will attempt to lay down on points in the complement of I(t) in L. Upon termination of the assignment of ones that begins at $[i,j; t_k]$, the random walk continues with no assignments made until at the (random) trial $\tilde{t}_{k+1} = t_k + \tilde{h}_{k+1}$, a lattice point $[i,j; \tilde{t}_{k+1}] \notin I(t_k) \cup J(t_k)$. A value $\xi_{t_{k+1}}$ of $\tilde{\xi}_{t_{k+1}}$ is generated, and the assignment of ones begins anew as described above.

Assignment of ones is governed by the following rules, where we let $N([i,j; t_k]) \equiv \{(i+x, j+y) \mid x = \pm 1,\ y = \pm 1\}$, the set of nearest neighbors to $[i,j; t_k]$ in L.

1. If no element of $N([i,j; t_k])$ is in $I(t_k)$, set $\delta([i,j; t_k]) = 1$.
2. If at least one element of $N([i,j; t_k])$ is in $I(t_k)$, set $\delta([i,j; t_k]) = 0$ and terminate the assignment of ones (from the "chain" of ξ_{t_k} ones).
3. If $\delta([i,j; t_k]) = 1$, let the random walk continue, repeating step 1 until either:

   (a) ξ_{t_k} ones have been assigned to $[i,j; t_k], \ldots, [i,j; t_k + \xi_{t_k}]$,

   or

(b) a position $[i,j;\ t_k + \ell]$, $\ell < \xi_{t_k}$, is reached for which at least one element of $N([i,j;\ t_k + \ell])$ is in $I(t_k)$.

Then terminate the assignment of ones from the "chain" of ξ_{t_k} ones.

Clearly, the random time $\tilde{h}_t = \tilde{t}_{k+1} - t_k$ depends upon the state S_{t_k} of the process at trial t_k.[4] And the number of ones assigned to lattice points from the "chain" $\tilde{\xi}_{t_k} = \xi_{t_k}$ of ones depends in a very complicated way on $S_{t_k}, S_{t_{k+1}}, \ldots, S_{t_{k+\ell}}$, where ℓ is the first integer such that $([i,j;\ t_{k+\ell}]) = 0$. In probabilistic parlance, the rules for generating a value of h_t and for the assignment of ones to lattice points are called stopping rules.

---

4. There is no semantic confusion in using "time" h_{k+1} to denote number of trials between $t_{k+1} - t_k$ and we shall do so.

Table 1

Summary List of Symbols

I. Model of the Discovery Process

A_i	surface area of i^{th} reservoir
k	number of wildcats drilled, i.e., number of prospects observed
M	number of prospective drilling sites
N	number of reservoirs in the basin
r	number of successful wildcats
R_M	total area of M prospects in the basin
s_k	total (cumulative) area of reservoirs discovered by k wildcats
S_N	total area of N reservoirs in the basin
$\underline{\theta}$	parameter set for the density function of A_i; $\underline{\theta} = (\mu,\sigma^2)$
u_k	total (cumulative) area of prospects drilled by k wildcats
x_i	outcome of i^{th} wildcat well ($x_i = 1$, where well is a success, 0 otherwise)

II. Spatial Model

$\delta(i,j)$	state of point (i,j); 0 or 1, where 1 signifies presence of petroleum
$I(t)$	petroleum areas: set of 1- points, $I(t) = [(i,j) \mid \delta(i,j) = 1]$
$J(t)$	nonpetroleum areas (or unassigned): set of 0- points, $J(t) = [(i,j) \mid \delta(i,j) = 0]$
L	spatial location: lattice of ordered pairs, $L = [(i,j) \mid i,j \text{ integer}]$
N	set of nearest neighbor points to point $(i,j;\ t_k)$: $N[(i,j;\ t_k)] \equiv [(i+x,\ j+y) \mid x = \pm 1,\ y = \pm 1]$
ξ_{t_k}	chain of ones laid down from point $(i,j;\ t_k)$ subject to prescribed stopping rules
S_t	state of the process at trial t: $S_t = [(i,j;\ t),\ I(t),\ J(t)]$

References

(1) M. Allais, "Method of Appraising Economic Prospects of Mining Exploration Over Large Territories", Management Science, July 1957, pp. 285-347.

(2) J. J. Arps and T. C. Roberts, "Economics of Drilling for Cretaceous Oil on the East Flank of Denver-Julesburg Basin", Bulletin of the American Association of Petroleum Geologists, Nov. 1958, pp. 2549-66.

(3) J. H. Engel, "Use of Clustering in Mineralogical and Other Surveys", Proceedings of the 1st International Conference on Operations Research, ORSA, Baltimore, 1957, pp. 176-92.

(4) S. M. Goldfeld, R. E. Quandt, and H. F. Trotter, "Maximization by Quadratic Hill-Climbing," Econometrica, July 1966, pp. 541-51.

(5) L. Golovin, "Two Mathematical Models for Oil and Gas Disposition," unpublished M.Sc. dissertation, Sloan School of Management, M.I.T., June, 1970, 65 pp.

(6) G. M. Kaufman, Statistical Decisions and Related Techniques in Oil and Gas Exploration (Englewood Cliffs, N.J.: Prentice-Hall, 1963), 307 pp.

(7) R. S. Uhler and P. G. Bradley, "A Stochastic Model for Determining the Economic Prospects of Petroleum Exploration Over Large Regions", Journal of the American Statistical Association, June 1970, pp. 623-30.