Mathematical Geology, Vol. 38, No. 2, February 2006 (© 2006)
DOI: 10.1007/s11004-005-9007-7

Evaluating Proportions of Undetected Geological Events in the Case of Erroneous Identifications[1]

Jerry L. Jensen,[2] Jeffrey D. Hart,[3] and Brian J. Willis[4]

Some geological events occur infrequently but still have a significant impact upon reservoir characteristics. By their very nature, however, it can be difficult to properly estimate the proportions of uncommon events because they may not appear during limited sampling. For example, even with 40 observations and an event proportion of 0.05, there is a 0.13 chance that no events will be observed. We provide some results and guidance concerning methods to estimate proportions when such events are not detected. Two cases are discussed, estimating proportions without errors in identification and estimating proportions when errors may arise.

It is well-known that the distribution of possible proportions in the error-free case can be calculated using Bayesian analysis. If one assumes a standard uniform distribution as the prior for the proportion, Bayesian analysis gives a Beta distribution for the posterior. The situation becomes more complicated, however, when detection errors are included; the true proportion has a distribution consisting of several Beta distributions. The difference in results between the error-free and with-error situations can be considerable. For example, when 10 error-free observations are made and no uncommon events are detected, there is a 0.50 chance that the true proportion exceeds 0.06 and a 0.10 chance that it exceeds 0.19. Including the effects of erroneous identifications, however, increases the median proportion to 0.09 and the upper decile to 0.27.

We also examine the case where there may be prior geological information, which can be incorporated by amending the prior distribution of the proportion. We find that the use of such a prior makes little difference unless there are very few observations or there are major differences between the anticipated and the observed proportions.

KEY WORDS: rare events; Binomial distribution; Beta distribution; erroneous detection; Bayes theorem.

INTRODUCTION

The ability of fluids to move through a hydrocarbon reservoir may reflect not only the mean rock permeability but also can depend critically on the abundance

[1] Received 6 December 2004; accepted 9 June 2005; Published online: 16 May 2006.

[2] Departments of Petroleum Engineering and Geology & Geophysics, Texas A&M University, 3116 TAMU, College Station, Texas, 77843-3116; e-mail:jensen@pe.tamu.edu.

[3] Department of Statistics, Texas A&M University, College Station, Texas, 77843-3143.

[4] Department of Geology and Geophysics, Texas A&M University, College Station, Texas, 77843-3115.

and lateral continuity of high- and low-permeability constituents within the rock volume. Where the rock is very heterogeneous, showing pronounced local variability in permeability, extreme-value permeability constituents may fundamentally control fluid movement, even where they comprise a small fraction of the rock volume.

Sensitivity of interwell- and field-scale reservoir permeability to infrequent extreme-value permeability rock constituents is observed in many geologic settings. For example, at smaller scales, sparse sandy burrow fills cutting a succession of otherwise continuous horizontal shale beds (Gingras and others, 1999), or dissolution vugs within an otherwise low-permeability carbonate (Lucia and Conti, 1987) can significantly increase the ability of fluids to move through the rock compared with that which would be predicted using average values. Similarly at larger scales, isolated high-permeability gravel lags concentrated along channel bases of fluvial deposits (Tye and others, 1999), formed during transgressive ravinement of shoreline sands (Caddel and Moslow, 2004) or as washed beds in debris-flow dominated fan deposits (Blair and McPherson, 1994), to give just a few examples, can result in sparse, isolated very high-permeability zones within stratigraphic units of significantly lower average permeability.

The sensitivity of reservoir behavior to infrequent extreme-value permeability rock constituents can be particularly pronounced in cases where the overall reservoir interval becomes more marginal and the mean rock constituents "tighter." For example, many of the lower Cretaceous units in western Canada are too quartz cemented to produce commercial quantities of gas where they comprise only sandstone, but units cut locally by even very-sparse, widely distributed "clean" gravels prove very economic under the right completion regime (Williams, Lerche, and Maubeuge, 1998).

It is critical to develop suitable methods to estimate the proportion of infrequent high-permeability constituents where they have paramount influence on reservoir behavior, even for cases where these constituents are rarely or never directly observed. Sensitivity of large-scale permeability to these infrequent events can be explained using percolation theory. In three-dimensional cases, only 10% or 12% of a particular rock type need to be present to satisfy the percolation threshold (Korvin, 1992, p. 22–23). Large clusters of a rock type can form at even smaller proportions. Numerical simulation studies (e.g., Desbarats, 1987) also show this effect.

A statistical procedure, based on the binomial distribution, can provide helpful error bounds on viable estimates of the proportions. Here, we review those results and extend them for the case where errors arise during the identification of these events, e.g., the situation where well logs may be used for facies identification but erroneous identifications are made.

ERROR-FREE EVENT IDENTIFICATION

In many reservoir characterization problems, uncertainties can be handled assuming they have a Gaussian (normal) distribution (e.g., Hurst and Rosvoll, 1991; Worthington, 2002). Proportions, however, may not be suited to this approach because values may be close to 0 or 1 and uncertainties based on the normal distribution may give non-physical values below 0 or more than 1. An approach using the binomial distribution gives better results.

The binomial distribution is well- known (e.g., Haas and Formery, 2002; Johnson, Kotz, and Kemp, 1993, p. 134–135) for describing the probabilities of outcomes in binary systems, assuming that outcomes are independent. If we have a system with two events, a and b, and make n measurements, the probability of observing type a, n_a times is given by the binomial distribution

$$\mathrm{prob}(n_a) = {}_nC_{n_a}(x)^{n_a}(1-x)^{(n-n_a)}, \tag{1}$$

where x is the true proportion of type a in the system, $0 \leq x \leq 1$, and ${}_nC_{n_a}$ is the number of ways that we can obtain n_a observations of type a out of n total observations. Without specifically assigning numbers for n, n_a, and x, we can extract two important results from Equation (1).

1. If $x = 0$, there is no chance we will have any observations of event a. This equation must therefore assume that we make no mistakes in identifying events. We treat this problem later.
2. If $x > 0$, there is a chance that we will not have any observations of a. That is,

$$\mathrm{prob}(n_a = 0) = (1-x)^{(n)}. \tag{2}$$

The second result is the one that suggests that we may have a problem when we want to identify infrequent but important events. For example, x may be small (e.g., 0.05) and n modest (say, 40), so that $\mathrm{prob}(n_a = 0) = 0.13$, which indicates that there is more than 1 chance in 10 that there will be no observations of a. Here is where we need some help when assigning appropriate values to proportions of these uncommon events in reservoir models.

From the analysis of well log, core, or other data, we can obtain values for n and n_a. With these two quantities and Bayes theorem (using a uniform prior), the Beta distribution gives the probability distribution of x, which we need for our

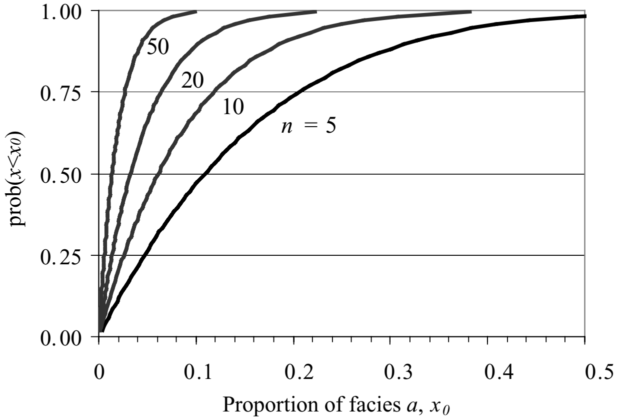

Figure 1. Integral of Equation (3) shows wide variation in possible value of the true proportion of event a, even when having observed no occurrences.

reservoir models (e.g., Lee, 1997, p. 77):

$$\text{prob}(x_0 \leq x < x_0 + \Delta x) = \frac{(n+1)!}{(n_a)!(n-n_a)!}(x_0)^{(n_a)}(1-x_0)^{(n-n_a)}\Delta x, \qquad (3)$$

where $0 \leq x_0 \leq 1$. The integral of Equation (3), the cumulative distribution function, shows especially well how n affects uncertainty about x. With $n = 50$ observations and $n_a = 0$ (no appearances of event a), there remains a 0.50 chance that the true proportion of type a is greater than 0.013 (Fig. 1). Smaller values of n give even larger median values of x. Thus, for example, if we have observed no occurrences of holes in a shale or fault, we still may want to include a small number of them in our reservoir models to evaluate their importance.

The above analysis assumes independence of the observations and a constant proportion x. These assumptions may not strictly hold for sampling of geological attributes (e.g., Haas and Formery, 2002). Johnson, Kotz, and Kemp (1993, p. 135) report, however, that the binomial model is fairly robust to mild departures and still gives useful results.

To this point, our Bayesian analysis has used a uniform prior distribution. The frequentist approach, which avoids prior probabilities altogether, gives a nearly identical result: the interval $[0, 1 - a^{1/n}]$ for a $100(1 - a)$ degree of confidence in x (Bickel and Doksum, 1977, p. 180–182). Later, we examine the sensitivity of results to the uniform prior.

EVENT IDENTIFICATION WITH ERRORS

We explicitly define two variables, E and F, in the event identification problem. Suppose that the true event identity is given by F, where $F = +1$ when type a is present and $F = -1$ when b is present. As in the error-free case, prob$(F = +1) = x$. Suppose also that a Bernoulli error variable E is present, and prob$(E = +1) = q$ and prob$(E = -1) = 1 - q$. E and F are assumed independent and their product, $G = EF$, defines another variable representing the apparent event observed at a particular location. When $E = +1$, $G = F$ and the apparent event is the same as the event which actually exists at a location. When $E = -1$, a mistake is made and the wrong event is observed. G is a Bernoulli variable with the probabilities

$$\text{prob}(G = +1) = xq + (1 - x)(1 - q) \tag{4a}$$

and

$$\text{prob}(G = -1) = x(1 - q) + (1 - x)q. \tag{4b}$$

When $q = 1$, we have the error-free case as discussed above.

Similar to the error-free case, using Equations (4) for n measurements of G we have

$$\text{prob}(n_a)_e = {}_nC_{n_a}[xq + (1 - x)(1 - q)]^{n_a}[x(1 - q) + (1 - x)q]^{(n-n_a)} \tag{5}$$

where the subscript e denotes error. Comparing Equations (1) and (5), the difference prob$(n_a = 0)_e -$ prob$(n_a = 0)$ is a function of both x and q, and may be either positive or negative. In the case of concern here, where x is small (e.g., less than 0.2) and assuming that our process of event identification is reasonably error-free, e.g., q exceeds 0.7 as in Kapur, Lake, and Sepehrnoori (2000), the difference is negative.

It is of interest to know what effect errors have on a common estimator of the proportion. In Bayesian analyses, a common point estimate is the mode of the posterior, which in our case of a uniform prior is simply the maximum likelihood estimate. It is easily verified that when q is known, the maximum likelihood estimate is

$$\hat{x} = \begin{cases} 0 & n_a \leq n(1 - q) \\ \dfrac{(n_a/n) - (1 - q)}{2q - 1} & n(1 - q) < n_a < nq \\ 1 & n_a \geq nq \end{cases} \tag{6}$$

where we are assuming that $q > 1/2$. In the error-free case, the maximum likelihood estimate is just the proportion n_a/n. It follows that for infrequent events, ignoring errors tends to *overestimate* the true proportion. This can be seen from the fact that $\hat{x} \leq n_a/n$ whenever $n_a/n \leq 1/2$. The only time the error-free and error-possible estimates agree (for $n_a/n \leq 1/2$) is when $n_a = 0$.

The overestimation phenomenon has the potential to be misleading. One might take it to mean that very small values of n_a provide greater than usual evidence that x is very small. Indeed this is not the case, as can be seen by examination of the posterior distribution of x. The example to follow and our application show that identification errors (when x is small) tend to *increase* the median and upper decile of the posterior distribution of x.

If q is not known exactly and we want to characterize it with a distribution, then the distribution of x becomes more complicated. Suppose we assume that q has a Beta distribution of the form

$$\operatorname{prob}(q_0 \leq q < q_0 + \Delta q) = \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)}(q)^{c-1}(1-q)^{d-1}\Delta q \tag{7}$$

where c and d are parameters that determine the shape and $\Gamma(\cdot)$ is the gamma function ($\Gamma(s+1) = s!$ for s a non-negative integer). For example $c = d = 1$ gives a uniform distribution for q, while $c = 2$ and $d = 1$ is a triangular distribution. Using Equations (5) and (7) with Bayes theorem and a uniform prior for x, we obtain the following *a posteriori* distribution for x when $n_a = 0$:

$$\begin{aligned}\operatorname{prob}(x_0 \leq x < x_0 + \Delta x) = {} & \frac{1}{r}\sum_{j=0}^{n} \frac{\Gamma(n+2)\Gamma(n+c-j)\Gamma(d+j)}{\Gamma(n-j+1)\Gamma(j+1)} \\ & \times (x_0)^{(j)}(1-x_0)^{(n-j)}\Delta x\end{aligned} \tag{8}$$

where

$$r = \sum_{j=0}^{n} \Gamma(n+c-j)\Gamma(d+j)$$

The distribution of x (Eq. 8) is a combination of Beta distributions. The values of c and d have an important effect on the result. Small values of c and d (e.g., $c = 2$ and $d = 1$) give a distinctly bimodal distribution with modes at $x = 0$ and $x = 1$ for any sample size. Larger values can still suffer bimodality, but the mode at $x = 1$ is much smaller for medium and large sample sizes ($n \geq 20$) and non-existent for small samples. The mode at $x = 1$ is a logical consequence of the case of a highly error-prone identification process which is showing no occurrences of event a.

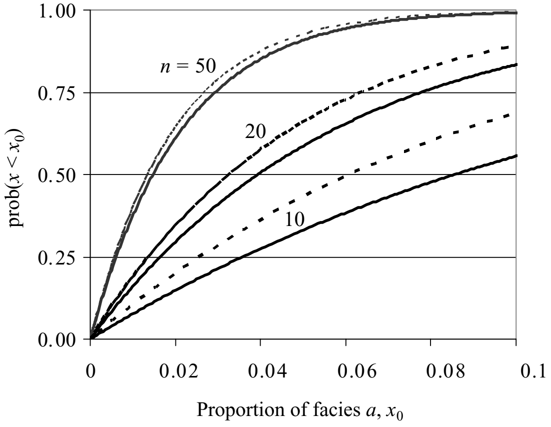

Figure 2. Integral of Equation (8) for the case $c = 13.3$ and $d = 4.44$ (solid lines) shows large variation in possible true proportion of event a, when having observed no occurrences and having a detection method which may make mistakes. Dashed lines are the error-free case.

Following Lee (1997, p. 79–80), we choose c and d to reflect experience about the accuracy of detecting event a by solving for c and d using

$$E(q) = c/(c + d) \tag{9a}$$

and

$$\mathrm{Var}\,(q) = (cd)/[(c + d)^2(c + d + 1)] \tag{9b}$$

where $E(\cdot)$ and $\mathrm{Var}(\cdot)$ are the expectation and variance, respectively. For example, using Equation (9) with $E(q) = 0.75$ and $\mathrm{Var}(q) = 0.01$ gives $c = 13.3$ and $d = 4.44$. As expected, comparing the cumulative distribution of x for the error-possible case with the error-free case shows a wider range of plausible values, particularly when n is 20 or less (Fig. 2). When $n = 10$, $c = 13.3$, and $d = 4.44$, for example, the error-free distribution of x has a median of 0.06 and a 90th percentile of 0.18 while the median and 90th percentile for the error-possible distribution of x is 0.086 and 0.27, respectively.

EFFECT OF PRIOR KNOWLEDGE UPON RESULTS

Bayesian analysis has often provided a powerful mechanism for combining data with prior knowledge (e.g., Kapur, Lake, and Sepehrnoori, 2000; Newendorp, 1972). Would prior information materially help the prediction of x? The derivation of Equation (2) uses Bayes theorem and assumes a uniform prior distribution for x (Lee, 1997, p. 77). That is, before we make any observations at the well, we assumed that

$$\text{prob}(x_0 \leq x < x_0 + \Delta x) = \Delta x.$$

Experience or other information could be incorporated by changing the prior distribution. If that distribution can also be expressed as a Beta distribution,

$$\text{prob}(x_0 \leq x < x_0 + \Delta x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(x_0)^{(\alpha-1)}(1-x_0)^{(\beta-1)}\Delta x$$

where α and β are parameters, then the posterior probability of x also has a Beta distribution

$$\text{prob}(x_0 \leq x < x_0 + \Delta x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+n_a)\Gamma(\beta+n-n_a)} \times (x_0)^{(\alpha+n_a-1)}(1-x_0)^{(\beta+n-n_a-1)}\Delta x. \tag{10}$$

For small n_a, the value of α can be quite influential on the posterior distribution. Thus, if we were expecting n_a to be large (e.g., $\alpha = 3$ and $\beta = 1$), this information would have a significant effect on the x distribution. Otherwise, where we are expecting x and n_a to be small, so that $\alpha < 2$, the resulting distribution for x will not be substantially changed by the choice of α and β. For example, if $\alpha = 1$ and $\beta = 2$ (a triangular distribution), there is little difference between the posterior distributions even for the $n = 5$ case (Fig. 3). These results agree with Lee (1997 , p. 87) and suggest that the results we give assuming a uniform prior are insensitive to this assumption.

APPLICATION

In a field study involving a fault block with only two wells, we evaluated well logs and observed facies a of proportion 0.15 at Well 1 but none at Well 2. The evaluation at Well 2 consisted of a meter-by-meter assessment over the reservoir interval of 42 m. Assuming our identification method made no errors, application of

Equation (3) when $n = 42$ shows prob($x < 0.05$) $= 0.89$ and prob($x < 0.016$) $=$ 0.50. Consequently, models with facies *a* proportions in the range 0–0.05 were used to evaluate the performance of the well and identify the range of uncertainties in the predictions. From several wells in other fault blocks, a detailed core analysis was available for comparison with the well log facies predictions, and we found the log evaluations of facies *a* were accurate 0.7 of the time with a standard deviation of 0.11. Using this information, Equation (8) shows prob($x < 0.06$) $= 0.90$ and prob($x < 0.018$) $= 0.50$, values which are close to the no-error case.

CONCLUSIONS

The case of determining the true proportion of small-probability (uncommon) events has been analyzed for the case where there may be identification errors. These events can affect reservoir performance and their presence may need to be accurately evaluated for useful reservoir models. We found that

- There is a significant probability that such events will not appear at a well, even with a large number of observations and no identification errors, while still being present at neighboring locations.
- The proportion of an event is a random variable with a Beta distribution, irrespective of sample size.
- Errors in event detection cause the unadjusted maximum likelihood estimates of the proportion of uncommon events to be too large.
- If the error rate for event detection is known, the posterior distribution of the proportion of uncommon events is a mixture of Beta distributions.

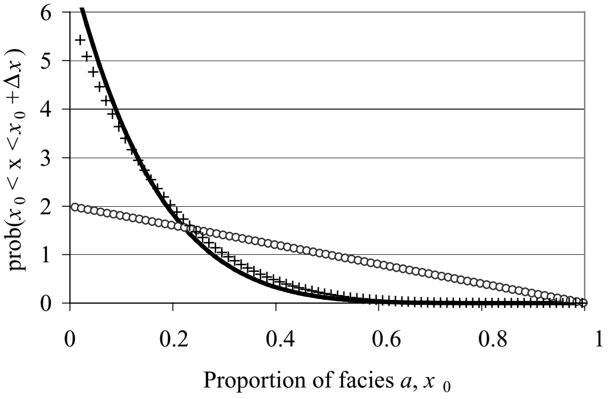

Figure 3. Triangular prior pdfs (open circles) and posterior pdfs using uniform (crosses) and triangular prior (black line) for the case $n = 5$.

- For surveys with detection errors and moderate or large sample numbers (i.e., 20 or more), the posterior distribution of the proportion is bimodal. Surveys with modest sample numbers (fewer than 20) may not be bimodal.
- Prior knowledge, or lack thereof, about the probability of an uncommon geological event can be modeled by a Beta distribution and then incorporated into a statistical analysis via Bayes theorem. This knowledge, however, has little effect on the posterior distribution unless the prior is very sharp.

REFERENCES

Bickel, P. J., and Doksum, K. A., 1977, Mathematical statistics: Prentice Hall, Englewood Cliffs, NJ, 493 p.

Blair, T., and McPherson, J., 1994, Alluvial fans and their natural distinction from rivers based on morphology, hydraulic processes, sedimentary processes, and facies assemblages: J. Sediment. Res., v. A64, no. 3, p. 450–489.

Caddel, E. M., and Moslow, T. F., 2004, Outcrop sedimentology and stratal architecture of the Lower Albian Falher C sub-Member, Spirit River Formation, Bullmoose Mountain, northeastern British Columbia: Bull. Can. Pet. Geol., v. 52, no. 1, p. 4–22.

Desbarats, A. J., 1987, Numerical estimation of effective permeability in sand-shale formations: Water Resour. Res., v. 23, no. 2, p. 273–286.

Gingras, M. K., Pemberton, S. G., Mendoza, C., and Henk, B., 1999, Assessing the anistropic permeability of Glossifungites surfaces: Pet. Geosci., v. 5, no. 4, p. 349–357.

Haas, A., and Formery, P., 2002, Uncertainties in facies proportion estimation I. Theoretical framework: The Dirichlet distribution: Math. Geol., v. 34, no. 6, p. 679–693.

Hurst, A., and Rosvoll, K. J., 1991, Permeability variations in sandstones and their relationship to sedimentary structures, *in* Lake, L. W., Carroll, H. B. , and Wesson, T. C., eds., Reservoir characterization II: Academic Press, New York, p. 166–196.

Johnson, N. L., Kotz, S., and Kemp, A. W., 1993, Univariate discrete distributions, 2nd ed.: John Wiley and Sons, New York, 565 p.

Kapur, L., Lake, L. W., and Sepehrnoori, K., 2000, Probability logs for facies classification: In Situ, v. 24, no. 1, p. 57–78.

Korvin, G., 1992, Fractal models in the earth sciences: Elsevier, Amsterdam, 396 p.

Lee, P. M., 1997, Bayesian statistics, 2nd ed.: Arnold, London, 344 p.

Lucia, F. J., and Conti, R. D., 1987, Rock fabric, permeability, and log relationships in an upward-shoaling, vuggy carbonate sequence: Bureau Economic Geology, University of Texas at Austin, Geological Circular 87–85, 22 p.

Newendorp, P. D., 1972, Bayesian analysis—A method for updating risk estimates: J. Pet. Technol., v. 24, no. 2, p. 193–198.

Tye, R. S., Bhattacharya, J. P., Lorsong, J. A., Sindelar, S. T., Knock, D. G., Puls, D. D., and Levinson, R. A., 1999, Geology and stratigraphy of fluvio-deltaic deposits in the Ivishak formation: Applications for development of Prudhoe Bay field Alaska: Am. Assoc. Pet. Geol. Bull., v. 83, no. 10, p. 1588–1623.

Williams, K. E., Lerche, I., and Maubeuge, F., 1998, Unconventional gas traps: Low permeability sands and gas accumulations: Energy Explor. Exploit., v. 16, p. 1–87.

Worthington, P., 2002, A validation criterion to optimize core sampling for the characterization of petrophysical facies: Petrophysics, v. 43, no. 6, p. 477–493.