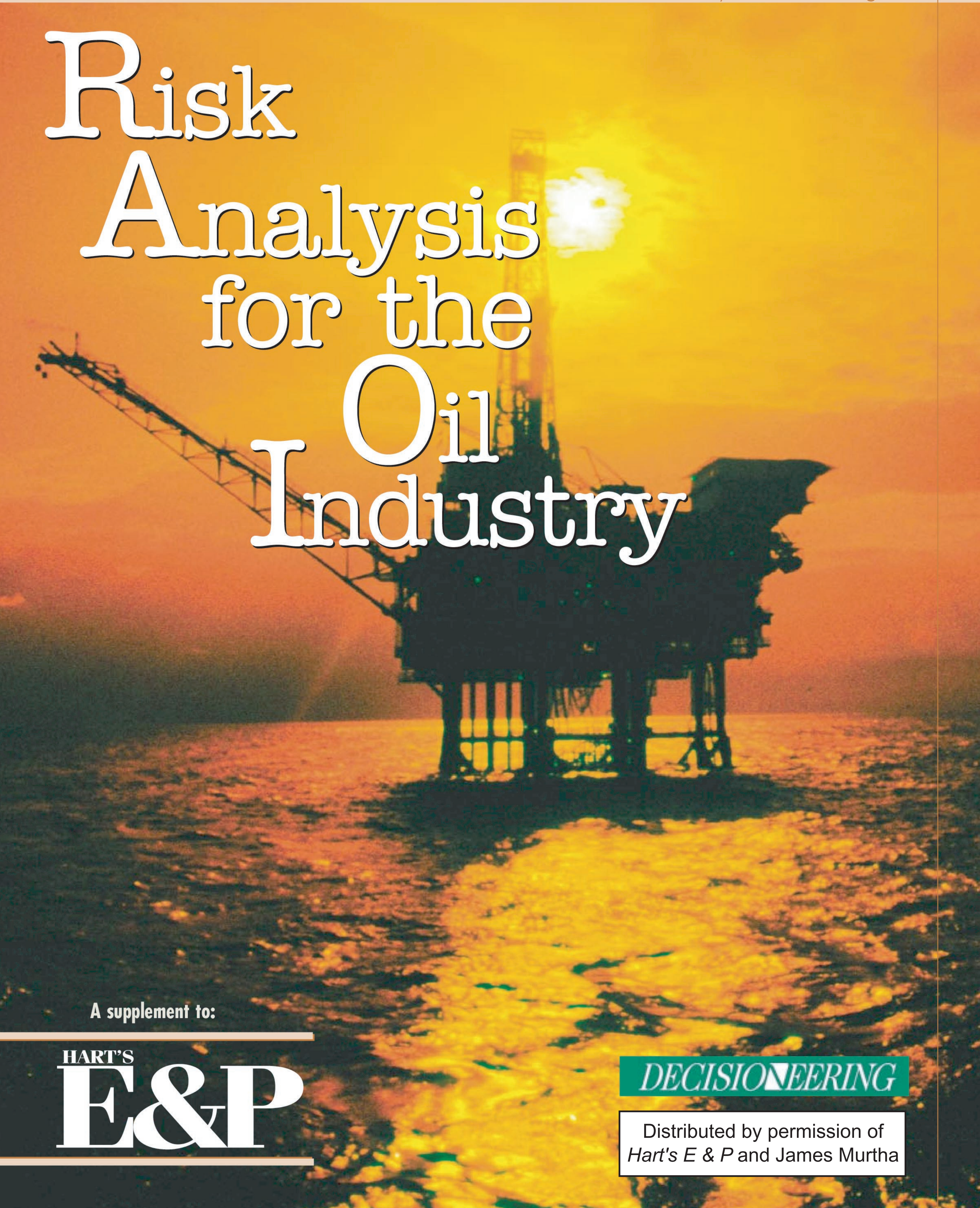
A Comprehensive
Study of Risk Management

Risk Analysis for the Oil Industry

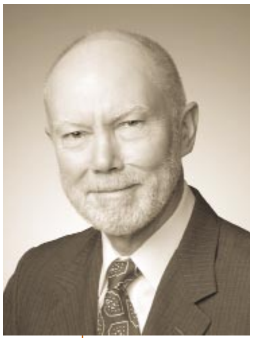


A supplement to:

HART'S E&P

DECISIONEERING

Biography



Jim Murtha, a registered petroleum engineer, presents seminars and training courses and advises clients in building probabilistic models in risk analysis and decision making. He was elected to Distinguished Membership in SPE in 1999, received the 1998 SPE Award in Economics and Evaluation, and was 1996-97 SPE Distinguished Lecturer in Risk and Decision Analysis. Since 1992, more than 2,500 professionals have taken his classes. He has published Decisions Involving Uncertainty – An @RISK Tutorial for the Petroleum Industry. In 25 years of academic experience, he chaired a math department, taught petroleum engineering, served as academic dean, and co-authored two texts in mathematics and statistics. Jim has a Ph.D. in mathematics from the University of Wisconsin, an MS in petroleum and natural gas engineering from Penn State and a BS in mathematics from Marietta College. ◆

Acknowledgements

When I was a struggling assistant professor of mathematics, I yearned for more ideas, for we were expected to write technical papers and suggest wonderful projects to graduate students. Now I have no students and no one is counting my publications. But, the ideas have been coming. Indeed, I find myself, like anyone who teaches classes to professionals, constantly stumbling on notions worth exploring.

The articles herein were generated during a few years and written mostly in about 6 months. A couple of related papers found their way into SPE meetings this year.

I thank the hundreds of people who listened and challenged and suggested during classes.

I owe a lot to Susan Peterson, John Trahan and Red White, friends with whom I argue and bounce ideas around from time to time.

Most of all, these articles benefited by the careful reading of one person, Wilton Adams, who has often assisted Susan and me in risk analysis classes. During the past year, he has been especially helpful in reviewing every word of the papers I wrote for SPE and for this publication. Among his talents are a well tuned ear and high standards for clarity. I wish to thank him for his generosity.

He also plays a mean keyboard, sings a good song and is a collaborator in a certain periodic culinary activity.

You should be so lucky. ◆

Table of Contents

When Does Correlation Matter?

What is correlation?

Often, the input variables to our Monte Carlo models are not independent of one another. For example, consider a model that estimates reserves by taking the product of area (A), average net pay (h) and recovery (R). In some environments, the structures with larger area would tend to have thicker pay. This property should be acknowledged when samples are selected from the distributions for A and h. Moreover, a database of analogues would reveal a pattern among the pairs of values A and h. Think of a cross-plot with a general trend of increasing h when A increases, as shown in Figure 1. Such a relationship between two variables can best be described by a correlation coefficient.

Examples of correlated variables

Typical pairs of correlated parameters are:

- daily high temperature and air conditioning cost for a personal residence in New Orleans;
- measured depth to total depth and drilling cost for wells in the North Sea;
- porosity and water saturation assigned to completion intervals for wells in a given reservoir;
- porosity and permeability assigned to completion intervals for wells in a given reservoir;
- operating cost and production rate;
- the permeability-thickness product and the recovery efficiency for gas wells in a given field;
- height and weight of 10-year old males in the United States;
- propane price and ethane price on Friday afternoon NYMEX;
- propane price and Brent crude price; and
- square footage and sales price of used houses sold in Houston in 1995.

Definition of correlation coefficient

Two variables are said to have correlation coefficient r if:

$$r = \frac{\mathrm{cov}(x, y)}{\sigma_x * \sigma_y}$$

where:

$$Cov = (1/n) \sum (x_i - \bar{x}) * (y_i - \bar{y})$$

$$\sigma_x = \sqrt{\mathrm{var}_x}$$

and

$$\mathrm{var}_x = (1/n) \sum (x_i - \bar{x})^2$$

Excel has a function, sumproduct ({x},{y}), that takes the sum of the products of the corresponding terms of two sequences {x} and {y}. Thus, covariance is a sumproduct.

The value of r lies between –1 (perfect negative correlation) and +1 (perfect positive correlation). Although there are tests for significance of the correlation coefficient, one of which we mention below, statistical significance is not the point of this discussion. Instead, we focus on the practical side, asking what difference it makes to the bottom line of a Monte Carlo model (e.g., estimating reserves or estimating cost of drilling a well) whether we include correlation. As we shall see, a correlation coefficient of 0.5 can make enough of a difference in some models to worry about it.

Before we illustrate the concept, we need to point out that there is an alternate definition.

Two types of correlation - ordinary and rank

When correlation was formally defined in the late 19th century, statisticians recognized one or a few points with extreme values could unduly influence the formula for calculating r. Specifically, the contribution of a pair (xi,yi) to the covariance, namely

$$\left(x_i - \bar{x}\right) * \left(y_i - \bar{y}\right)$$

could be an order of magnitude larger than the other terms.

Charles Spearman introduced an alternative formulation, which he labeled “distribution-free” and called it the rank-order correlation coefficient, in contrast to the Pearson coefficient defined above. To obtain the Spearman coefficient, one replaces the original data with their ranks then calculates the correlation coefficient using the ranks. Tables 1 and 2 provide a simple example illustrating both methods.

Impact of correlation on the output of a model

What does correlation do to the bottom line? Does it alter the distribution of reserves or cost or net present value (NPV), which is, after all, the objective of the model? If so, how?

We can make some generalizations, but remember Oliver Wendell Holmes's admonition, "No generalization is worth a damn…including this one."

First, a positive correlation between two inputs will result in more pairs of two large values and more pairs of two small values. If those variables are multiplied together in the model, for example, a reserves model, then this results in more extreme values of the output.

Even in a summation or aggregation model (aggregating production from different wells or fields, aggregating reserves, estimating total cost by summing line items and estimating total time), positive correlation between two summands will cause the output to be more disperse.

In short, in either a product or aggregation model, a positive correlation between two pairs of variables will increase the standard variation of the output.

The surprising thing is what happens to the mean value of the output when correlation is included in the model.

For product models, positive correlation between factors will increase the mean value of the output. For aggregation models, the mean value of the output is not affected by correlation among the summands.

Let us hasten to add that many models are neither pure products nor pure sums, but rather complex algebraic combinations of the various inputs.

Example 1. Correlating parameters in a volumetric reserves model

A simple example will illustrate the properties of a product model.

A common classroom exercise to illustrate the effect of correlation on reserves uses the model:

N = AhR

With:

A = Triangular (1000,2000,4000)
h = Triangular (20,50,100)
R = Triangular (80,120,200)

Table 1. Area, porosity, and gas saturation (and their ranks) for 13 reefal structures

Name	Area, km^2	Porosity	Sg	Rank_Area	Rank_por	Rank_Sg
S1	10	0.12	0.77	3	1	3
M1	24	0.12	0.85	6	1	10
A1	37	0.13	0.87	9	3	11
P2	6	0.14	0.81	1	4	6
K3	28.8	0.14	0.91	7	5	13
D	6	0.15	0.61	1	6	1
U1	34	0.15	0.82	8	7	8
P1	13	0.16	0.81	5	8	6
K2	60	0.16	0.78	12	9	4
K4	11	0.17	0.83	4	10	9
K1	58	0.18	0.80	11	11	5
U2	48	0.22	0.75	10	12	2
Z1	108	0.22	0.89	13	12	12

Table 2. Correlation and Rank Correlation coefficients for the reefal structure data.

Ordinary r	Area	Porosity	Sg	Rank r	Area	Porosity	Sg
Area	1			Area	1		
Porosity	0.67	1		Porosity	0.57	1	
Sg	0.36	-0.02	1	Sg	0.28	-0.12	1

Each student is assigned a correlation coefficient, from 0 to 0.7, and uses it to correlate A with h and h with r (the same coefficient for both pairs). The students run the simulations using both these correlations and call out their results, which are tallied in Table Y:

r	mean	StDev
0	17.5	7.7
.1	17.9	8.6
.2	18.1	9.1
.3	18.3	9.7
.4	18.6	10.5
.5	18.8	10.9
.6	19.0	11.3
.7	19.3	11.9

Now of course, even if both correlations were warranted, it would be unlikely they would be identical. Nevertheless, we must note the standard deviation can increase by as much as 50% and the mean by about 10% under substantial correlation.

The message is clear: pay attention to correlations in volumetric product models if you care about the danger of easily underestimating the mean value by 5% or more, or if you care about understating the inherent uncertainty in the prospect by 30% or 40%.

Almost as important: often correlation makes little difference in the results. We just need to check it out.

While it may not be obvious, nearly all the correlations in reserves models cause the dispersion (and the mean) to increase. It is far less common for the correlations to have the opposite effect.

In more complex models, however, it is possible for correlations to actually reduce uncertainty. One example we discovered several years ago is a capital project involving construction and operation of a gas-fired electric plant. The high positive correlation between price of natural gas (the principal operating cost) and price of electricity (the principal factor in revenue) will cause the output, NPV, to be far more modest in range (on the order of 50%) when the correlation is included, making the investment less risky.

With a cost model, adding correlations between line items is one of several refinements that can increase the range of total costs. Correlation coefficients in this case are often thought to be higher than those in reserve models. For instance, steel prices can influence numerous line items, thereby resulting in a positive correlation among the pairs. One of our clients used a cost model with a matrix of risk factors vs. line items. When the same risk factor influenced two or more line items, these items became positively correlated.

Positive correlation (the usual type) among line items will cause the total cost to have a bigger range, but the mean cost is unaffected by correlation of any kind.

It should be noted that correlation can have a serious impact on cost models in terms of contingency. Some people define contingency for the total project as the difference between some large percentile, say P85 or P90, and a base value like P50 or the mean total cost. This contingency could easily grow substantially, though, because when we correlate line items, the mean (and to a large degree the P50) remains unchanged but the dispersion (typically) increases.

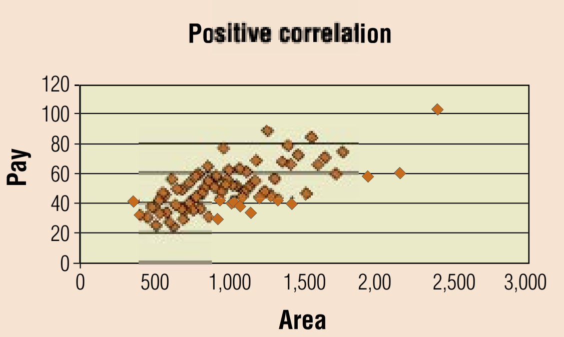


Figure 1. Correlation between Area and Pay

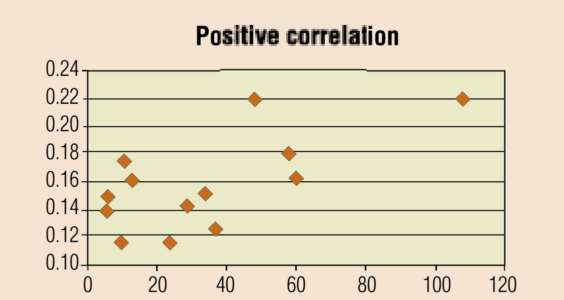


Figure 2. Original data, r = 0.67

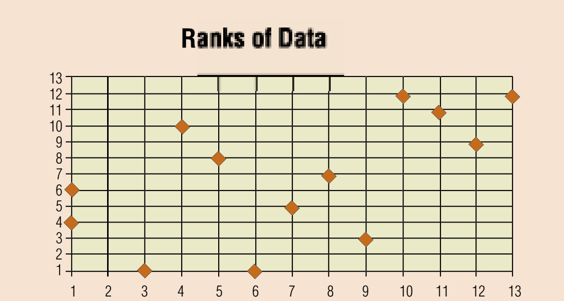


Figure 3. Ranks of data r = 0.57

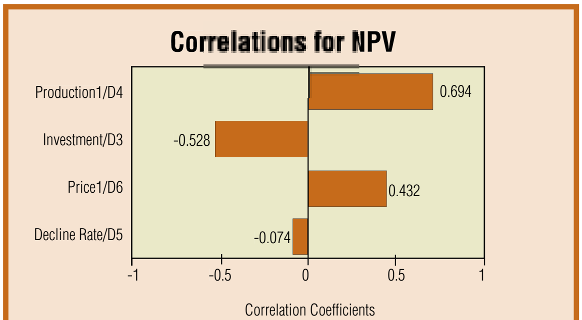


Figure 4. Sensitivity chart

Estimating the correlation coefficient

Once you decide to include correlation in a model, you must find a way to estimate the coefficients. This issue is similar to the more basic issue of what distribution to use for a given input variable. In both cases, you can resort to one or more of the following:

- empirical data;
- experience; and
- fundamental principles.

Sometimes we have adequate and appropriate field data, in which case, we can calculate the (rank order) correlation coefficient. Likewise, we can use curve-fitting procedures (both popular Monte Carlo software packages have them) to generate the probability distribution.

If we have no data or have appropriate data but only a small sample, we should temper our automatic processes with experience. What does this kind of parameter usually look like: is it symmetric or skewed, small or large? And for correlation, how do these two parameters usually relate to each other: positively or negatively?

Once in a great while, we can appeal to that higher power: fundamental principles. For example:

- "this parameter is a sum or average of other parameters; thus it should be approximately normal;"
- "this parameter is a product of other parameters – it should be skewed right;"
- "these two prices represent substitutable commodities, therefore they should be positively correlated;" and
- "these two parameters represent competing quantities and should be negatively correlated."

Remember, the correlation coefficient is limited to the range [–1,+1]. Anything in the range [–0.15, +0.15] could be noise and is unlikely to impact the results. It is rare to have correlation coefficients above 0.8 or below –0.8. So, make an educated guess, and run the model with several alternate values and document the results. You may not resolve the issue, but you can defuse it.

Another use of correlation: Sensitivity

One argument for studying correlation is to better understand sensitivity analysis. Figure 4 shows one common type of sensitivity chart, sometimes called a tornado chart. In this case, the numbers represent rank correlation coefficients between the target output variable (in case a model has two or more outputs) and each of the inputs. A negative correlation coefficient indicates a model input that tends to increase as the output decreases and vice versa. Thus, as capital gets large, NPV tends to get small. The relationship is properly described by correlation, because the other three inputs also affect NPV. Simply knowing that capital is large does not force NPV to have a certain small value. But given a particular, large value of capital might limit the range of NPV to a smallish part of the entire NPV range, a fact made clear by Figure 5.

Summary

Things to remember:

- correlation is easy to calculate in Excel (the function Correl);
- there are two types of correlation coefficients: ordinary (Pearson) and rank-order (Spearman). They tend to differ when one or both the variables is highly skewed;
- correlation might matter; it depends on the type of model and the strength of the correlation;
- correlation always will affect the standard deviation, often but not always increase it;
- correlation will affect the mean value of a product; and
- correlation is useful to describe sensitivity of output to inputs. ◆

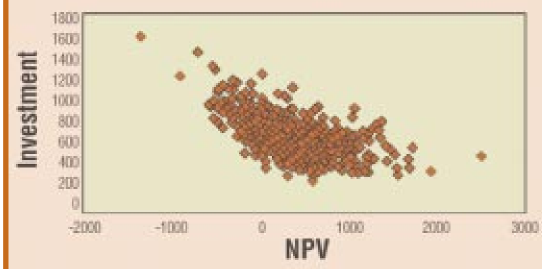


Figure 5. Negative correlation: investment and NPV