Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

USING COPULAS IN RISK ANALYSIS

Dalton F. Andrade
Pedro A. Barbetta
Paulo J. de Freitas Filho
Ney A. de Mello Zunino

Federal University of Santa Catarina
University Campus – Trindade
Florianópolis , SC, 88040-900, BRAZIL

Carlos Magno C. Jacinto

CENPES – Well Technology Engineering
PETROBRAS S.A.
Avenida Hum, Quadra 7, Ilha do Fundão
Rio de Janeiro, RJ, 21941-598, BRAZIL

ABSTRACT

Nearly every well installation process nowadays relies on some sort of risk assessment study, given the high costs involved. Those studies focus mostly on estimating the total time required by the well drilling and completion operations, as a way to predict the final costs. Among the different techniques employed, the Monte Carlo simulation currently stands out as the preferred method. One relevant aspect which is frequently left out from simulation models is the dependence relationship among the processes under consideration. That omission can have a serious impact on the results of risk assessment and, consequently, on the conclusions drawn from them. In general, practitioners do not incorporate the dependence information because that is not always an easy task. This paper intends to show how Copula functions may be used as a tool to build correlation-aware Monte Carlo simulation models.

1 INTRODUCTION

The total time taken in drilling and completion operations of oil and gas wells are subject to considerable uncertainty and risk factors, due to the limited knowledge concerning the geologic characteristics of the formation, technical difficulties and unexpected behavior of human operators (Jacinto 2002). Moreover, this time represents 70 to 80% of the final cost of the well due to the high costs of daily rent of the drilling and completion rigs. The planning and risk assessment of these activities are hindered by unexpected events such as kicks (bags of gas), loss of circulation and well collapse. Those events can cause waste of time, increasing costs, decline of the production or even the loss of the well (Jacinto 2002).

Risk analysis and management of petroleum exploration ventures is growing worldwide and many international petroleum companies have improved their exploration performance by using principles of risk analysis in combination with new technologies (Harbaugh 1995, Rose 2001).

Lately, Monte Carlo simulation has been the preferred technique for many well forecasting applications related to operational risk analysis. The total time taken in well drilling and completion operations and their associated cost are two good examples of variables considered in such forecasting.

In this study we discuss some problems involved in the estimation of the total time taken in well drilling and completion operations and present one solution that makes use of copulas, mathematical functions that allow for the generation of joint distributions of dependent random variables. The next section presents a short description of oil and gas well engineering and tries to identify the uncertainty and risk factors present in the well accomplishment. A brief description of the Monte Carlo tool and some of the problems in its utilization are presented in Section 3. In Section 4, we show how not considering dependence when simulating dependent data can affect the total time and, in Section 5, we introduce and discuss copulas as one tool to generate dependent data. Finally, in Section 6, we show one application example of the proposed methodology before we present our final remarks in Section 7.

2 DRILLING AND COMPLETION ENGINEERING AND RISK ANALYSIS

2.1 Drilling and Completion Operations

The development of a petroleum field includes many activities: drilling and completion of wells, installation of fluid collector systems (manifolds and flexible lines), construction and installation of a production unity (petroleum platform), installation of the production drain flow system (oil and gas pipelines, oil ships) (Jacinto 2002).

The drilling of an oil well is accomplished through a rig. The rocks are drilled by the action of the rotation and weight applied to an existent drill in the extremity of a drilling column. Rock fragments are continually removed through a drilling fluid or mud, which is pumped into the interior of the drilling column by an injection head (swivel) and comes back to the surface through the ring space formed between the walls of the well and the column. When a certain depth is reached, the column is removed and a coating column goes down into the well. The space between the coating tubes and the walls of the well is cemented with the purpose of isolating the crossed rocks, allowing the progress of the drilling. In this way, the well is drilled in several phases, characterized by the different diameters of the bits (Jacinto 2002).

When the drilling is finished, a new stage of operations, designed to prepare the well so it can produce in safe and economic conditions during its useful life is carried out: the completion. In this phase, the valves in the head of the well that control the flow of petroleum are installed. The well is conditioned and shelled, and the production column is installed. Then the production of petroleum can begin (Jacinto 2002).

2.2 Risk Analysis

Risk connotes the possibility of loss and the chance or probability of that loss. Modern risk analysis utilizes principles of statistics, probability theory and utility theory (Jain 1991, Bedford 2001, Vose 2001). In oil exploration there are many aspects of risk. Risk and uncertainty are associated with drilling operations, with field development and with production. In this paper, we are going to concentrate on those elements of risk associated to the drilling and completion of individual wells (Jacinto 2002). If the operations needed to drill and complete a given well are carried out without problems, the total time is usually short. On the other hand, if the same well has a few setbacks, failures, accidents and even if workovers occur (such as equipment failure, drill breaks, wall tumbling or a well blowout), the total time could be much longer than expected. So, when forecasting the total time, it must be expressed by a probability distribution, instead of a single number.

The components of well drilling and completion times are often difficult to define with any degree of accuracy or exactitude and the failure sources can be blunder, systematic or random, associated with operation, equipment, material, geology or workmanship (Harbaugh 1995).

3 THE TRADITIONAL MONTE CARLO METHOD FOR RISK EVALUATION

Because of the probabilistic nature associated with the time of drilling and completion operations, the estimation of the necessary time to rent all the required rigs is considered a complex task. The scenario where the analyst takes decisions is full of uncertainties for nearly every action. Therefore, several of them are risky decisions.

One of the most traditional techniques to deal with decision and risk analysis under uncertainty is modeling and simulation using the Monte Carlo method. Considering the assumption that the analyst can associate a theoretical random distribution, which better describes each operation in the process, it is possible to model and simulate the system by random sampling from the input distributions. In this case, the defined functions are related to the time to conclude each drilling and completion operation. In its great majority, these are random variables (Law 1991, Jain 1991, Bedford 2001, Vose 2001, Evans 2002, Coelho 2005).

During this research, we developed a customized simulation tool (E&P Risk) that allows the estimation of the total time necessary to execute all needed operations. Before performing the simulation, the analyst should define the representative distribution for each operation. In the E&P Risk suite, this can be done by searching the operation time from the corporate database and performing a fitting process using a built-in tool. For every operation, an input distribution can be adopted and fed into the model.

In the first version of the tool, we assumed that all the operations were independent. At the end of the simulation, after generating hundreds or even thousands of samplings of the operation time, an estimation of the total time is presented in conjunction with a risk exposition histogram, with the indication of some desired percentiles to better support the decisions (Figure 1). As the histogram and their related results (estimated total time and cost) are presented, the decision maker can now use those values to take a decision and/or use them to refine it after confronting it with those obtained with the aid of complementary approaches like the one we are going to explain in the next topic.

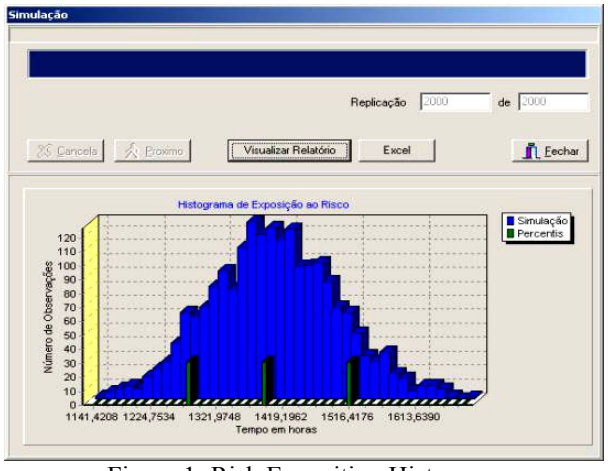

Figure 1: Risk Exposition Histogram

One problem related to the use of the Monte Carlo method is the independence assumption for all the input distributions. If the completion times of one or more opera-

tions are actually correlated, the implications on the final results of the simulation could be remarkable. In such cases, the simulation results might even be considered unsuitable for decision making.

4 IMPLICATIONS IN CONSIDERING INDEPENDENT, THINGS THAT ARE DEPENDENT

Let us consider T_1, T_2, …, T_K random variables that represent the times involved in the execution of each of a given set of operations (e.g. some well drilling and completion operations, as mentioned above). Then, it is well known that the mean or expected value and the variance of the total time $T = T_1 + T_2 + \ldots + T_K$ are given by

$$E(T) = \sum_{k=1}^{K} E(T_k)$$

and

$$Var(T) = \sum_{k=1}^{K} Var(T_k) + 2\sum_{k=1}^{K}\sum_{k'<k}^{K-1} Cov(T_k, T_{k'})$$

where $E(T_k)$ and $Var(T_k)$ are, respectively, the mean value and the variance of each T_k, and $Cov(T_k, T_{k'})$ is the covariance between T_k and $T_{k'}$. The latter is a measure of dependence between two random values and it is equal to zero when they are independent. Therefore, unless all the covariances between the random variables are zero, the variance of T will not be equal to the sum of the individual variances. As the covariance can be either positive or negative, by considering independent, things that are dependent, one can end up with a value for the variance of the total time quite different, either below or above, from the true value.

As we can see, dependence does not affect the mean total time, but it can very badly affect its variance, which will affect the extreme values of the total time distribution, such as P10 and P90, for example. Table 1 shows that via statistics for three values of ρ, Pearson's correlation coefficient. (X and Y are normal random variables with means 30 and 40, and variances 25 e 64, respectively.)

Table 1: Statistics Related to T = X + Y for Three Different Values of ρ

Type of Dependence	P10	P25	Mean P50	P75	P90	Variance
ρ = 0 (independence)	57.9	63.6	70	76.4	82.1	89
ρ = 0,8 (positive dependence)	54.1	61.7	70	78.3	85.9	153
ρ = -0,8 (negative dependence)	63.6	66.6	70	73.4	76.4	25

Therefore, it is very important to consider dependence when generating bivariate, or higher dimension, dependent data.

5 GENERATING MULTIDIMENSIONAL DATA - COPULAS

In order to generate multidimensional data, we need to know the joint distribution of the random variables. For dimension 2, $H(x,y) = Prob(X \le x, Y \le y)$. Considering independence, $H(x,y) = F_1(x)F_2(y)$, where $F_1(x)$ and $F_2(y)$ are the marginal (one-dimensional) distributions of X and Y, respectively. So, in the independence case, it is sufficient to know the two marginal distributions to construct the joint distribution. When we have dependence, besides that we will need also to know the type and extent of the dependence.

One probabilistic model which has been considered in the literature to accomplish that is the multivariate normal distribution, where the dependence is well defined by Pearson's correlation coefficient, ρ. When we are modeling the time of drilling and completion operations, we do not expect that the normal distribution will be appropriate to model each individual time of the operations, because they usually follow asymmetric distribution. Therefore, by considering asymmetric distributions we would be more realistic. In the literature, we can find some multivariate asymmetric distributions, like, for instance, the Gumbel's bivariate exponential distribution (Nelsen, 1999) and multivariate gamma distributions (Mathai and Moschopoulos, 1991) that could be used. However, they have two drawbacks for our purposes. They require the same marginal distributions and also, the interpretation of the dependence between the random variables is not, in general, easy to understand. In this work, we propose the use of copulas to generate joint distributions.

Copulas are functions that associate a point in the unit square [0,1]x[0,1], for the two-dimensional case, to a point in the interval [0,1]. Formally, Copula is any function $C : [0,1]^n \rightarrow [0,1]$ with the following properties:

1. $C(u_1,u_2,...,u_n)$ is increasing in each argument $u_i \in [0,1]$, i=1,2,...,n;
2. $C(1,1,...,u_i,1,...,1) = u_i$ for all i;
3. For all $(a_1,a_2,...,a_n)$, $(b_1,b_2,...,b_n) \in [0,1]^n$ with $a_i \le b_i$, we have

$$\sum_{i_1=1}^{2}\cdots\sum_{i_n=1}^{2}(-1)^{i_1+\ldots+i_n} C(u_{1i_1},\ldots u_{ni_n}) \ge 0,$$

with $u_{j1} = a_j$ and $u_{j2} = b_j$, j=1,...,n.

One example of a two-dimensional, one-parameter copula is the following function:

$$C_\theta(u,v)=\left(-\frac{1}{\theta}\right)\cdot\ln\left\{1+\frac{\left[e^{-\theta u}-1\right]\cdot\left[e^{-\theta v}-1\right]}{e^{-\theta}-1}\right\}$$

with $\theta\in\Re\setminus\{0\}$, the copula parameter. This copula is known as Frank Copula. Examples of copulas can be found in (Nelsen 1999), (Xue-Kun Song 2000), and in many other publications.

Now, if we consider u = F_1(x) and v = F_2(y), where F_1 and F_2 are any two one-dimensional distributions, associated to two random variables X and Y, then one joint distribution of X and Y with a dependence parameter θ is

$$H(x,y)=C_\theta(F_1(x),F_2(y))=$$
$$=\left(-\frac{1}{\theta}\right)\cdot\ln\left\{1+\frac{\left[e^{-\theta F_1(x)}-1\right]\cdot\left[e^{-\theta F_2(y)}-1\right]}{e^{-\theta}-1}\right\}.$$

Therefore, using this approach we can construct joint distributions from any one-dimensional distributions. For instance, if F_1 is the cumulative distribution function of a triangular distribution, with parameters $\alpha < \beta < \gamma$, and F_2 is the cumulative distribution function of an exponential distribution, with mean μ, then

$$H(x,y)=\begin{cases}0, & \text{for } x<\alpha \text{ and } y<0,\\ -\frac{1}{\theta}\ln\left\{1+\frac{[e^{\frac{\theta(x-\alpha)^2}{(\gamma-\alpha)(\beta-\alpha)}}-1]\cdot[e^{-\theta(1-e^{-\frac{y}{\mu}})}-1]}{e^{-\theta}-1}\right\}, & \text{for } \alpha\le x<\beta \text{ and } y\ge 0,\\ -\frac{1}{\theta}\ln\left\{1+\frac{[e^{-\theta\left(1-\frac{(\gamma-x)^2}{(\gamma-\alpha)(\gamma-\beta)}\right)}-1]\cdot[e^{-\theta\left(1-e^{-\frac{y}{\mu}}\right)}-1]}{e^{-\theta}-1}\right\}, & \text{for } \beta\le x<\gamma \text{ and } y\ge 0,\\ 1-e^{-\frac{y}{\mu}}, & \text{for } x\ge\gamma \text{ and } y\ge 0.\end{cases}$$

All the information about the dependence between the two random variables is in the parameter θ, whose value can be interpreted in terms of Kendall's τ measure of association. In our context, it is important to have the dependence between the random variables expressed in terms of a coefficient like Kendall's τ, because it is not affected by strictly increasing transformations of the random variables. It assumes values in the interval [-1, 1], being negative when the two variables are negatively correlated, and being positive when they are positively correlated. The closer τ is to 1 (or -1), the stronger will the dependence between the two variables be. The zero value means no correlation. It can be shown that τ and θ are related to each other by the following expression:

$$\tau=\tau(X,Y)=4\cdot\int_0^1\int_0^1 C_\theta(u,v)\cdot\left[\frac{\partial^2 C_\theta(u,v)}{\partial u\partial v}du\,dv\right]-1$$

For the Frank Copula, we have that (see Embrechts, 2001)

$$\tau=1-\frac{4\cdot[1-D(\theta)]}{\theta},\ \text{where}\ D(\theta)=\frac{1}{\theta}\int_0^\theta\frac{t}{e^t-1}dt$$

Another advantage of copulas is that there are algorithms that can be implemented to generate dependent data. Below we present one algorithm for our Frank Copula setup:

1. Generate *u* e *w* from two independent U[0, 1] distributions;
2. Evaluate, for a given θ (or τ),

$$v=C_{V|U}^{-1}(u,w)=\frac{1}{\theta}\cdot\ln\left\{\frac{w.e^{-\theta}+e^{-\theta u}-w\cdot e^{-\theta u}}{e^{-\theta u}-w\cdot e^{-\theta u}+w}\right\};$$

3. Evaluate $x=F_1^{-1}(u)$ and $y=F_2^{-1}(v)$. The pair (*x*, *y*) is a pair of random numbers with dependence defined by θ (or τ);
4. Repeat 1–3 as many times as it is the desired number of pairs.

In the next section we present one application example using simulated data.

6 APPLICATION EXAMPLE

We now introduce a case study in order to illustrate the significance of observing the correlations among random variables when doing risk analysis. The case study is comprised of a series of experiments which deal with two supposedly correlated well operations. We are interested in estimating the total time (and, consequently, the cost) for the completion of those operations. For the purposes of this exercise, let us assume that the completion times for the first operation (operation A) are given by a triangular distribution and that those for the second (operation B) are

given by an exponential one. Table 2 shows the parameters of the two distributions.

Table 2: Parameters of the Marginal Distributions

Operation	**Distribution for Completion Time**
A	Triangular(216, 288, 456)
B	Exponential(150)

The dependence between the operations is expressed by means of the Frank Copula, which is used to produce the bivariate data for the experiment. The two distributions shown above are the marginal distributions used by the copula function. Another parameter, τ (Kendall's correlation coefficient), specifies the degree of dependence between two generated values of a pair.

To illustrate the effects of considering the correlations when doing bivariate generation of random variables, the experiment will be repeated with different values for the dependence parameter of the copula, including negative dependence. The experiments, with 10,000 replications each, generate a value for the first and second completion times as described above in section 5. Table 3 shows the results for P10, P50 and P90, percentiles obtained for different values of τ (Kendall's Correlation Coefficient).

Table 3: Percentiles for Different Values of τ

Exp.#	Correl τ.	Percentiles			
		P10	P50	P90	P90 - P10
1	-0,95	399	425	602	203
2	-0,80	389	429	607	218
3	-0,60	371	433	620	249
4	-0,40	351	437	641	290
5	-0,20	331	434	645	315
6	None	314	434	676	362
7	0,20	300	429	695	395
8	0,40	290	427	713	423
9	0,60	280	420	724	444
10	0,80	275	419	738	463
11	0,95	273	422	742	469

In the experiment number 6 the values were generated in an independent way, meaning that the second completion time is not correlated with the first one. If we look at the results of that experiment, we observe that the distance from P10 to P90 is about 362 hours. The P50 – P90 gap is larger than the P10 – P50, as a result of the skewed Triangular and Exponential distributions. Consider now the results of experiments 7 to 11. As the correlation coefficient increases, so does the distance between P10 and P90. In experiment number 10, for instance, the simulated time for operation B is correlated to that of operation A, according to a coefficient of 0.8. In this case, the distance from P10 to P90 is now about 463 hours. At this level of correlation and taking only two operations into account, the difference for the total time could be of more than four days. On the other hand, looking now at the experiments 1 to 5, the results are the opposite. As the coefficient negatively increases, the gap between P10 and P90 reduces. If again we analyze in terms of risk assessment, the same reasoning applies, i.e., not considering a negative correlation could result in a super estimation for the total time. And, in this case, if we remember that the daily rent of the drilling and completion rigs in petroleum wells could cost more than US$250,000.00 a day, these differences are very significant. Figure 2 shows the evolution of the P10, P50 and P90 percentiles as a function of the Kendall's correlation coefficient.

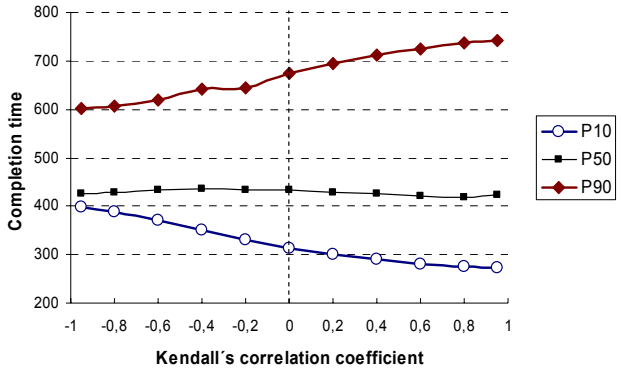

Figure 2: Percentiles as a Function of the Correlation Coefficient

7 FINAL REMARKS

Risk assessment is an important constituent in the development process of a well installation. Well drilling and completion operations, especially in deep waters, are very risky and uncertain operations, subject to great variability. That variability influences the final cost severely, making the employment of more elaborate techniques a factor of most importance.

One of the most traditional techniques to deal with decision and risk analysis under uncertainty, especially when dealing with well forecasting, is the application of probabilistic methods, in particular Monte Carlo Simulation.

Associated with Monte Carlo simulation models, copula functions provide a simple, yet powerful framework that allows for the appreciation of the dependence among correlated operations, while not imposing restrictions on the marginal distributions used to model them.

When empirical data related to the processes involved are available, practitioners can verify which copula would better fit their data. That can be done by using, for instance, maximum likelihood based methods. Details can be found in Mendes and Melo (2005).

REFERENCES

Bedford, T., and R. Cooke. 2001. Probabilistic Risk Analysis: Foundations and Methods. Cambridge University Press.

Bishop, C. M. 1995. Neural Networks for Pattern Recognition. Oxford University. U.K.

Coelho, D. K., M. Roisenberg, P. J. Freitas, and C. M. Jacinto. Risk Assessment of Drilling and Completion Operations in Petroleum Wells Using a Monte Carlo and a Neural Network Approach. In *Proceedings of the 2004 Winter Simulation Conference*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Embrechts, P., F. Lindskog, and A. McNeil. 2001. Modeling dependence with copulas and applications for risk management. Department of Mathematics, ETHZ, Switzerland. <http://www.risklab.ch/ftp/papers/DependenceWithCopulas.pdf>.

Evans, J. R., and D. Olson. 2002. Introduction to Simulation and Risk Analysis. 2nd Edition, Prentice Hall.

Jacinto, C. M. C. 2002. Modeling and Risk Analysis of Drilling and Completion operations of Deep Waters Oil and Gas Wells M.S. Thesis (in Portuguese). UFF, Rio de Janeiro, Brazil.

Jain, R. 1991. The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling. John Wiley & Sons Inc, New York, U.S.A.

Harbaugh, J. W., J. C. Davis, and Wendebourg. 1995. Computing Risk for Oil Prospects: Principles and Programs. Pergamon. U.K.

Haykin, S. 1998. Neural Networks: a Comprehensive Foundation. McMillan College. U.S.A.

Law, A. M. Simulation Modeling and Analysis. 2nd Edition. McGraw-Hill, New York, U.S.A.

Mathai, A.M., and P. G. Moschopoulos. 1991. On a Multivariate Gamma. *Journal of Multivariate Analysis*, 39, pp. 135-153.

Mendes, B. V. M., and E. F. L. Melo. 2005. Robust fit for copulas model. *Proceedings of the Second Brazilian Conference on Statistical Modeling in Insurance and Finance*. USP, São Paulo, Brazil. pp. 32-44.

Nelsen, Roger B. 1999. An Introduction to Copulas. New York, NY, U.S.A. Springer-Verlag.

Nelsen, Roger B. 2005. Dependence modeling with arquimedian copulas. *Proceedings of the Second Brazilian Conference on Statistical Modeling in Insurance and Finance*. USP, São Paulo, SP, Brazil. pp. 45-54.

Rose, P. 2001. Risk Analysis and Management of Petroleum Exploration Ventures. *AAPG Methods in Exploration Series*. No. 12. AAPG. U.S.A.

Vose, D. 2001. Risk Analysis, A Quantitative Guide. 2nd Edition. John Willey & Sons.

Xue-Kun Song, P. 2000. Multivariate dispersion models generated from Gaussian copula. Scandinavian Journal of Statistics Vol. 27. pp. 305-330.

AUTHOR BIOGRAPHIES

DALTON F. ANDRADE is professor in the Department of Computer Science at Federal University of Santa Catarina (UFSC), Brazil. His research interests include statistical methods for large educational assessment, longitudinal data analysis and dependence modeling.. He is a member of AERA – American Educational Research Association, IASI – Inter American Statistical Institute and ABE – Brazilian Statistical Association . His e-mail address is <dandrade@inf.ufsc.br>.

PEDRO ALBERTO BARBETTA is an associate professor of the Department of Informatics and Statistics at the Federal University of Santa Catarina – Brazil. He received a Ph. D. in Industrial Engineer, in 1998, from the same university. He is the author of two textbook in statistics education. Professor Barbetta is a member of the Brazilian Statistical Association. His areas of interest are Design of Experiments and Multivariate Data Analysis. His e-mail address is <barbetta@inf.ufsc.br>.

PAULO J. FREITAS FILHO is an associate professor in the Department of Computer Science at the Federal University of Santa Catarina (UFSC), Brazil. His research interests include simulation of computer systems for performance improvement, risk modeling and simulation, analysis for input modeling and output analysis. He is a member of SCS – Society for Computer Simulation, SBC – Brazilian Society for Computers. His e-mail address is <freitas@inf.ufsc.br>.

NEY A. DE MELLO ZUNINO is an undergraduate student at the Federal University of Santa Catarina, in the final term to get a bachelor's degree in Computer Science. He has been involved with computing from an early age, having worked in diverse areas ranging from computer graphics animation to knowledge management systems. As of 2005, he's joined the Performance Lab. at UFSC, working along a team of researchers in the implementation and test of simulations with Copula-related algorithms. His e-mail address is <zunino@inf.ufsc.br>.

CARLOS MAGNO C. JACINTO has been a Petrobras S.A. (Brazilian Energy Company) employee for nearly 17 years. He is a doctoral candidate in COPPE–UFRJ. His research interests include risk modeling and simulation, performance improvement, failure prediction, and artificial intelligence, all applied to Well Technology Engineering. His e-mail address is <cmcj@petrobras.com.br>.