MODELING AND SIMULATION OF PROCESS INTERRUPTIONS WITH COPULAS

Ivan Ourdev[(a)], Simaan AbouRizk[(b)]

[(a,b)]Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada

[(a)]iourdev@ualberta.ca, [(b)]abourizk@ualberta.ca

ABSTRACT

Construction projects are exposed to various unexpected interruptions, with equipment breakdowns being the most common. Proper modeling of these interruptions along with the associated uncertainty can significantly reduce risk and improve project management. We employ a new mathematical approach, the copula method, to model the field observations in a tunnel excavation project. We characterize the excavation process interruptions with their degree of severity and frequency, and develop a Student t copula model for the underlying dependence structure. The model is consecutively utilized in a Monte Carlo simulation that incorporates all available information to forecast the project completion time. The adaptive estimates then serve as a basis for project management decisions. Our approach allows to incorporate the uncertainty within an intuitive simulation framework and to accurately model the dependence of the different dimensions of the process interruptions.

Keywords: copulas, uncertainty, process interruptions, Monte Carlo simulation, project management

1. INTRODUCTION

Uncertainty permeates real-life project management: uncertain durations, uncertain cost, sudden weather changes, equipment breakdown, human resource problems, unexpected changes in project scope, etc. Uncertainty is rarely beneficial and takes the form of a risk that must be dealt with. It threatens the bottom line and, particularly, the project schedule. Many project activities are sequential, and alterations to the duration of some tasks have a ripple effect on the start times of all subsequent tasks down the activity chain. Although a certain amount of contingency time is normally built into all project schedules, changes in the schedule have to be managed in a timely fashion in order to ensure a relatively smooth flow of labor and materials. Thus, the forecasting of task execution times becomes an essential ingredient of successful project risk management.

The common approach to decisions made under uncertainty relies on probability theory, where the quantities of interest are considered random variables (r. v.) described by probability distributions. The mean of the probability distribution specifies the expected value of the modeled quantity, while the standard deviation quantifies our uncertainty about the 'true' value of this mean. The most common distribution used by researchers and practitioners alike is the normal, also called Gaussian, distribution, which takes the familiar bell shape. The popularity of this distribution is due to its convenient mathematical properties. It is analytically tractable, completely described by only two parameters: the mean and the standard deviation. A linear combination of normal distributions is also a normal distribution with parameters determined by the means and the covariance matrix of the original components. Also, according to the central limit theorem the distribution of the sum of many independent r. v. with finite variances approaches normal distribution. However, in real life, normal distributions are exceptional, rather than the rule.

The standard approach to modeling data generated by a vector-valued random process is to fit a multivariate probability distribution, using e.g. the maximum likelihood (ML) method. The drawbacks of this approach are the lack of control in the fitting process and imprecise physical interpretation of the components of the distribution.

Recently, new mathematical objects, called copulas, have become very popular for multivariate modeling of dependent variables and risk management (Nelsen 2003, Yan 2006, Frees and Valdez 1998). Copulas are multivariate functions with uniform marginals that allow the construction of the joint distribution from the constituent marginals capturing the dependency structure of the latter (see e.g. Nelsen 2006). The intuitive copula approach to building multivariate distributions is a two-step statistical procedure. In the first step, the empirical marginals are obtained by fitting univariate distributions to the data. In the second step, an appropriately chosen copula is used to combine the univariate marginals into a joint distribution. It has been pointed out (Mikosch 2006) that the use of this approach is not universally justified, but for our purposes it has definite advantages.

The first advantage of the two-step methodology is that it enables meaningful interpretations of the marginal distributions, and it applies to and compares with existing, well-researched models. The second advantage is that, by choosing a specific copula, we can tailor the fitting process to the relative importance of the dependence domain.

In this work, we model the unexpected process interruptions in a tunnel excavation project. The model was included as a separate component of a much larger distributed decision support and planning system, based on discrete-event simulation. The purpose of the model was to serve as basis for Monte Carlo simulations, where the randomly generated interruptions of the excavation are taken into account for an adaptive project schedule planning.

Using copulas allowed us to borrow the familiar intuition from actuarial science, where the unexpected losses are characterized by two random variables: severity and frequency (Klugman 2004). In our case, these become the marginal distributions of the severity of the interruptions and the time intervals between interruptions. The separate choice of the copula allows the importance of the relatively rare but severe breakdowns to be stressed.

The paper is organized as follows: Section 2 introduces the notion of copulas and presents the main definitions and important properties. It also includes some broad examples of copulas and gives accounts of the methodology for statistical inference and simulation. This paper considers a case with two random variables, so for the sake of simplicity and notational clarity we only use bivariate copulas, but all the results presented are also valid in higher dimensions (see e.g. Nelsen 2006 for a general treatment). Section 3 contains an overview of the tunnel excavation operations and the data collection, particularly of process interruptions. The copula model of the excavation process interruptions and the results from the simulations are presented in Section 4, which also contains a brief description of the software framework that encompasses the model. The conclusion, Section 5 contains an evaluation of the approach and some suggestions for future research. Some well-known probability concepts are included in Appendix A to serve as an easy reference for comparing the properties of two-dimensional probability distributions and copulas.

2. COPULAS

2.1. Definitions for bivariate copulas

The notion of mathematical copulas was introduced by Abe Sklar in 1959 as functions that link n--dimensional distributions to their one--dimensional margins (Sklar 1959). Copulas are, in general, distribution functions that have as arguments $\mathbb{R}^m$ – valued random vectors $\mathbf{X} = \left(X_1, \cdots, X_m\right)$; for simplicity, we restrict this presentation to the two-dimensional case, $\mathbb{R}^2$. Formally, in the case of two r. v., X and Y, the (bivariate) copula, C, is a function $[0,1]^2 \to [0,1]$ that has the following properties:

- It is a grounded function:

$$C\left(u,0\right) = C\left(0,v\right) = 0, \quad \forall u,v \in [0,1] \qquad (1)$$

- It is consistent with its margins:

$$C\left(u,1\right) = u, \quad C\left(1,v\right) = v, \quad \forall u,v \in [0,1] \qquad (2)$$

- It has a non-negative C – volume, i.e. for every $u_1, u_2, v_1, v_2 \in [0,1]$, such that $u_1 \le u_2$, and $v_1 \le v_2$ the following inequality holds:

$$C\left(u_2,v_2\right) - C\left(u_2,v_1\right) - C\left(u_1,v_2\right) + C\left(u_1,v_1\right) \ge 0 \qquad (3)$$

This property requires copulas to be "2-increasing" functions, which is the two-dimensional analog of a nondecreasing function of one variable (see Nelsen 2006 for details).

The basis for the theory of copulas is Sklar's theorem, which states that for a two-dimensional joint cdf $F_{X,Y}(x,y)$ with marginal distributions $F_X(x)$, $F_Y(y)$ there exists a unique 2-copula such that:

$$F_{X,Y}(x,y) = C\left(F_X(x), F_Y(y)\right) = C\left(u,v\right) \qquad (4)$$

If the random variables X and Y are continuous, then Equation 4 is unique. Otherwise, the copula is uniquely determined on the range $Ran(X) \times Ran(Y)$. Conversely, if C is a bivariate copula and $F_X(x)$, $F_Y(y)$ are distribution functions, then the function $F_{X,Y}(x,y)$ defined by Equation 4 represents a joint cdf. Thus, copula "couples" the marginals to form a joint cdf.

One of the methods for copula construction is to use Sklar's theorem and invert the expression of Equation 4 as

$$C\left(u,v\right) = F_{X,Y}\left(F_X^{-1}(u), F_Y^{-1}(v)\right) \qquad (5)$$

For continuous r. v., copulas, as every ordinary joint cdf, have their corresponding densities, *c*, analogously to Equation A2:

$$c\left(u,v\right) = \frac{\partial C\left(u,v\right)}{\partial u\, \partial v} \qquad (6)$$

and the bivariate joint pdf has the following canonical representation:

$$f_{X,Y}(x,y) = c\left(F_X(x), F_Y(y)\right) f_X(x) f_Y(y) \qquad (7)$$

This representation illustrates the decomposition of the multivariate probability density, $f_{X,Y}(x,y)$ into one-dimensional marginals, $f_X(x)$, $f_Y(y)$, and a dependence structure specified by the copula density, C.

2.2. Examples of copulas

Given that every multivariate distribution has a corresponding copula, the number of possible copulas is enormous. There are three copula families that have been found most useful: elliptical, Archimedean, and extreme-value copulas.

Elliptical copulas derive from elliptical distributions, with the two main representatives, Gaussian and Student's distributions. They are widely used for modeling financial time series, particularly in the context of factor models (Malevergne and Sornette 2005). The parameter that is needed for their specification is the correlation matrix in the multivariate case, or the correlation coefficient, ρ, in the bivariate. Figure 1 shows the copula densities for two elliptic copulas with the same correlation coefficient $\rho = .5$ a normal copula and a Student t copula with 3 degrees of freedom.

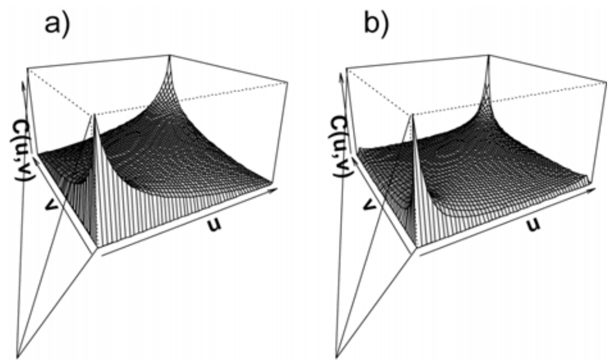

Figure 1: The Copula Densities for Two Elliptic Copulas with a Correlation Coefficient ρ = 0.5: (a) Normal Copula, and (b) Student t Copula with 3 Degrees of Freedom.

Student t copulas are particularly interesting because of their higher densities in the corners $(1,0)$, and $(0,1)$, as seen on Figure 1b. We illustrate the effect of the correlation coefficient on the density distribution of a Student copula that links two beta marginal distributions. Beta distribution is a flexible distribution with density

$$b(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{8}$$

where α, and β are shape parameters and the normalization constant is the beta function, $B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$.

Figure 2 shows an example of two marginal beta distributions with parameters $f_X(x;3,1.5)$, and $f_Y(y;2,5)$. The copula used to link these marginals is the Student t with two degrees of freedom. Figure 3 shows the contour plots for different values of the correlation coefficient, ρ.

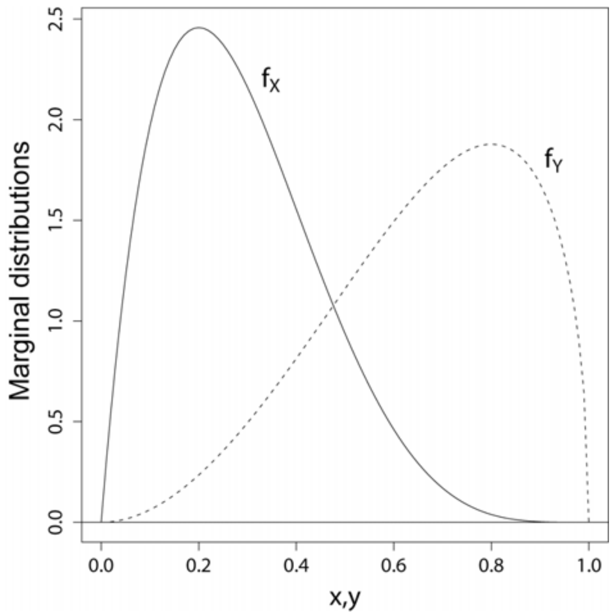

Figure 2: Two Marginal Beta Distributions with Parameters $f_X(x;3,1.5)$ and $f_Y(y;2,5)$

Archimedean copulas often arise in the context of the actuarial modeling of sources of risk (Frees and Valdez 1998). They are constructed without referring to the distribution functions. Instead, the construction is done using a continuous strictly decreasing function, $\varphi:[0,1]\to[0,\infty]$, called generator, such that $\varphi(1)=0$, and the formula:

$$C(u,v) = \varphi^{[-1]}(\varphi(u)+\varphi(v)) \tag{8}$$

Where $\varphi^{[-1]}$ is the pseudo-inverse of φ, defined as:

$$\varphi^{[-1]} = \begin{cases} \varphi(u), & \text{if} \quad 0 \le u \le \varphi(0), \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

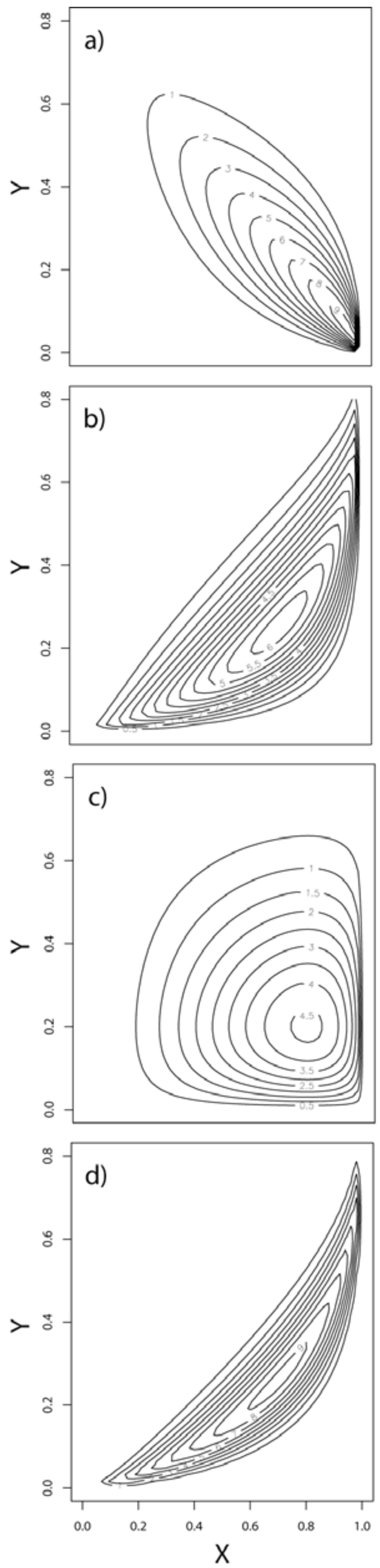

Figure 3: Contour Plot of Student t Copulas with Two Degrees of Freedom that Link the Beta Marginals from Figure 2 with Different Correlation Coeffieicents: a) $\rho = -.75$, b) $\rho = .75$, c) $\rho = 0$, and d) $\rho = .9$

Different generators give rise to different copulas. For example, the Clayton copula

$$C_\theta^{\mathrm{Cl}}(u,v) = \max\left(\left[u^{-\theta} + v^{-\theta} - 1\right]^{-1/\theta}, 0\right) \quad \forall\theta \in [-1,\infty), \tag{11}$$

is obtained from the generator:

$$\varphi_\theta^{\mathrm{Cl}}(u) = \frac{1}{\theta}\max\left(u^{-\theta} - 1\right) \tag{12}$$

Figure 4 shows the contour plots of the Clayton copula with parameter $\theta = 1$ and its density. It is clear that Clayton copula has a heavy lower tail, the $(0,0)$– corner. This copula is very important in the multivariate statistics of extremes, because it can be shown that it is the limiting copula for the class of the Archimedean copulas when the probability level of the quantiles approaches zero.

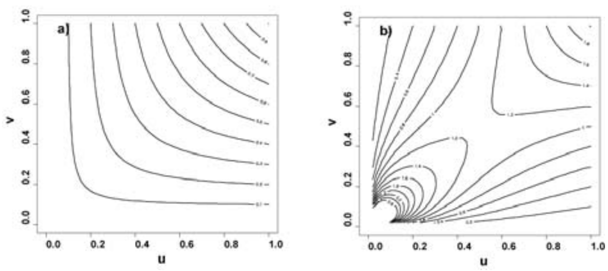

Figure 1: Contour Plots of the Clayton Copula with Parameter $\theta = 1$, Panel (a), and its Density, Panel (b).

Another popular example is the Gumbel copula

$$C_\theta^{\mathrm{G}}(u,v) = \exp\left(-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{-1/\theta}\right) \quad \forall\theta \in [1,\infty), \tag{13}$$

which is obtained from the generator:

$$\varphi_\theta^{\mathrm{G}}(u) = (-\ln u)^\theta \tag{14}$$

The Gumbel copula has a density that peaks at the $(1,1)$ corner and is, in a sense, complementary to Clayton copula which density is the highest at the $(0,0)$ corner. The Gumbel copula is also an example of extreme-value copulas, which are derived from generalized extreme value distributions.

2.3. Statistical inference and simulation

The majority of the copula estimation approaches rely on the maximum likelihood technique. ML can be applied either to the joint estimation of the parameters of the marginals and the copula, or the two can be treated separately. We follow the latter approach, because it gives a better control on the estimation process. The method, called inference function for margin (IFM) (Joe and Xu 1996), is a two-step procedure: first, the marginals are fitted to the data, and then the copula is estimated conditionally on the fitted marginals. Both the fitting of the marginals and the copula involve a choice of the appropriate distributions. Fitting univariate distributions to data is a well-studied problem (see e. g. Joe and Xu 1996). The identification of the appropriate copula is still largely empirical. Currently, there is a non-parametric identification methodology only for the Archimedean copula class (Genest and Rivest 1993, Wang and Wells 2000).

The best fit for the marginals was found to be the gamma distributions. The probability density function of the gamma distribution, as parameterized by the shape parameter, α, and the rate parameter, λ, is given by

$$g(x;\alpha,\lambda)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} \tag{15}$$

where the normalization constant is the gamma function, $\Gamma(\alpha)=\int_0^{\infty}t^{\alpha-1}e^{-\alpha}dt$. The expression $X\sim Gamma(\alpha_X,\lambda_X)$ is used to signify that the random variable X has a gamma distribution with the corresponding shape and rate parameters.

The fitting of the marginal distributions is done in two steps. First, using the method of the moments we obtain rough estimates for the shape parameter, $\hat{\alpha}$, and the rate parameter, $\hat{\lambda}$. Then we use the moment estimates as a starting point for the maximum-likelihood estimation step.

The method of moments is a well-known technique for obtaining parameter estimates. The construction is done by matching the sample moments, m_i, with the corresponding distribution moments μ_i, and solving for the latter. In the case of the gamma distribution only the first two sample moments are needed, $m_1=\overline{x}$, and $m_2=(n-1)s^2/2$, where $\overline{x}$ is the usual sample mean, s^2 is the sample variance and n is the sample size. The corresponding (central) distribution moments are defined as $\mu_i=E\left(X^i\right)$ and for the gamma density, Equation 15, the integrations give a mean $\mu_1=\mu=\alpha/\lambda$, and a variance $\mu_2=\sigma^2=\alpha(\alpha+1)/\lambda^2$. The matching step yields the following starting estimates:

$$\hat{\lambda}_0=\frac{n\overline{x}}{(n-1)s^2-n\overline{x}^2},\qquad \hat{\alpha}_0=\frac{n\overline{x}^2}{(n-1)s^2-n\overline{x}^2}. \tag{16}$$

These initial estimates are used as starting points for the ML step, which maximizes the sample likelihood.

The IFM method for a two-parameter marginal distribution functions, as in our case, involves a clear separation of the marginal parameters $\boldsymbol{\beta}=(\alpha,\lambda)$ from the association parameters $\boldsymbol{\theta}$. The likelihood function for n independent observations $\{x_k:k=1,\cdots,n\}$ and a density distribution g is defined as

$$L(x_1,\cdots,x_n;\boldsymbol{\beta})=\prod_{k=1}^{n}g(x_k;\boldsymbol{\beta}) \tag{17}$$

Substitution of the gamma density, Equation 15, in this expression yields the following form of the log-likelihood function:

$$l(\boldsymbol{\beta})=n\alpha\log\lambda-n\log\Gamma(\alpha)+(\alpha+1)\sum_{k=1}^{n}\log x_k-\lambda\sum_{k=1}^{n}x_k \tag{18}$$

For our bivariate case, there are n pairs of observations of interruptions $\{(x_k,y_k):k=1,\cdots,n\}$ with severity, x_k, occurring at intervals, y_k. We also need to introduce additional index, M, that enumerates the two marginals, $M=X,Y$. Thus the IFM estimates for the parameters $\hat{\boldsymbol{\beta}}_M$ of two gamma distributions with densities, g_X, and g_Y, become

$$\hat{\boldsymbol{\beta}}_M=\arg\max_{\boldsymbol{\beta}_M}\sum_{k=1}^{n}\log g_M\left(m_k;\boldsymbol{\beta}_M\right),\quad M=X,Y \tag{19}$$

The second step finds the IFM estimate of the association parameter, $\boldsymbol{\theta}$, using the copula density, c, as

$$\hat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}\sum_{k=1}^{n}\log c\left(F_X\left(x_k;\boldsymbol{\beta}_X\right),F_Y\left(y_k;\boldsymbol{\beta}_Y\right)\right) \tag{20}$$

The ML estimates, obtained above form the foundation of Monte Carlo simulations (Fishman 2003, McLeish 2005). The simulation of process interruptions modeled by a specific copula uses Sklar's theorem. The approach relies on some algorithm (e.g. Malevergne and Sornette 2005) for generation of two uniform random numbers u_X, and u_Y, on the interval $(0,1)$ with a dependence structure given by the copula, C. In order to generate two random variables, X, and Y, from the proper joint distribution $F_{X,Y}(x,y)=C\left(F_X^{-1}(x),F_Y^{-1}(y)\right)$, only the application of the generalized inverse is needed: $x=F_X^{-1}(u_X)$, and $y=F_Y^{-1}(u_Y)$.

We apply the steps for modeling and simulation described above to the data for excavation process interruptions.

3. DATA

The data consists of the durations of the delays and interruptions in the stage SW3 of the South Edmonton Sanitary Sewer (SESS) tunneling project in the City of Edmonton, Canada. The project involves the excavation of a 3.5 km long sanitary sewer tunnel using a tunnel boring machine (TBM). It started in February 2006 and was completed in August 2007. The tunneling operations are constantly monitored and the relevant data is recorded and collected by a decision support system, called COSYE.

The information about the daily operation of the TBM comes from two sources: one is an engineering survey system called TACS (tunnel advance control system), and the other is the report prepared at the end of the day. The daily report contains information about the number of work shifts per day, the length of the shifts in hours, and the source and the duration of the project delays and interruptions.

3.1. TBM operations

Tunnel construction by means of tunnel boring machines is considered to be a state-of-the-art technology. The main TBM element is a cylindrical rotating cutterhead with a diameter approximately equal to that of the tunnel that bores in the earth strata. The support for the forward press is provided by gripper shoes that engage outwardly with the tunnel wall. The support for the wall of the tunnel is provided by one-meter long cement rings that are placed as the tunnel is being dug. Each ring consist of two semi-circular segments, called liners. The front part of the tunnel, where the actual excavation takes place is called the tunnel face.

The SW3 tunnel has a relatively small diameter, 2.34 m, which to a large extent determines the tunneling operations. It has a single-track railway for most of the tunnel length, which becomes a double-track only in the area close to the entrance shaft. Still, in order to save time on loading and unloading operations, two trains carry loads between the face of the tunnel and the entrance shaft.

The excavation is a batch process with activities naturally partitioned into cycles. The beginning of a cycle is marked by the unloading of the liners from the train. The unloaded train is positioned behind the TBM and excavation begins. The carts of the train collect the dirt from the excavation. After the one-meter length is excavated, the train, loaded with dirt, starts traveling back towards the entrance shaft, while the TBM begins the installation of the liner blocks. The loaded train dumps the dirt into a sump pocket, while the first train, already loaded with liner blocks, starts traveling towards the face of the tunnel. The crane hoists the dirt from the sump pocket to the surface, where it is stockpiled. Afterward, the crane lowers down the liners blocks for the next segment of the tunnel. This completes one cycle of tunnel operations.

3.2. Excavation interruptions

Tunnel construction is a process of several interdependent activities, placing the equipment under a significant strain. Machine breakdowns and system malfunctions are common and often result in interruptions of the whole chain of operations. In our approach, we disregard the specific source of interruption. No distinction between a breakdown in the excavation process and an interruption in some of the support operations is made. The system of tunneling operations is modeled as a whole. The main reason for this approach is that we do not have enough data to model the elements separately. Another reason is the high degree of coupling (correlation) between the system elements. The general system approach solves both problems.

Thus, the only assumption we make is that the characteristics of the system will remain practically constant for the duration of the project until completion. For example, no new equipment will be introduced, or the load on the existing equipment will remain the same. The soil composition profile of the site, which is one of the main determinants of the load on the TBM, has little variation as inferred on the basis of the exploratory borehole samples. Also, experience indicates little effect due to seasonal changes.

Figure 5 shows the interruptions that occurred between September 14, 2006 and May 10, 2007. The x-axis represents the time line measured in work shift operating hours. The breakdown's severity is measured in terms of the work shift time it takes to fix the problem. The excavation operations take place during work-shifts. Normally there is one 10-hour shift per day, with weekends off, but depending on the overall progress of the project or external events, the project manager can decide on splitting the work into unusual 8-hour shifts, one or two per work-day. Only the time during which the equipment is in operation contributes to the probability of breakdowns, thus only work shift time is taken into account. The frequency of the interruptions is quantified by the interval between two successive breakdowns. The severity of the break is measured in terms of the work shift time it takes to fix the problems and restart excavation.

4. MODEL

4.1. The COSYE system

The model of the process interruptions was embedded in the general simulation and decision support system COSYE (Construction Synthetic Environment). Details of the system are given elsewhere (AbouRizk and Mohammed 2000). Here we only include a very brief general description.

The COSYE simulation environment is a .NET implementation of the HLA (High Level Architecture) IEEE standard for modeling and simulation (SISC 2000). The HLA architecture is a general framework for creating complex distributed simulations from relatively independent simulation units called federates. It has two main elements: the federate interface specification (FIS), and the object model template (OMT). FIS specifies the communication interface for combining the individual simulation components and maintaining the interoperability between them, while OMT describes the exchanged data. The model execution is provided by a run time infrastructure (RTI) server.

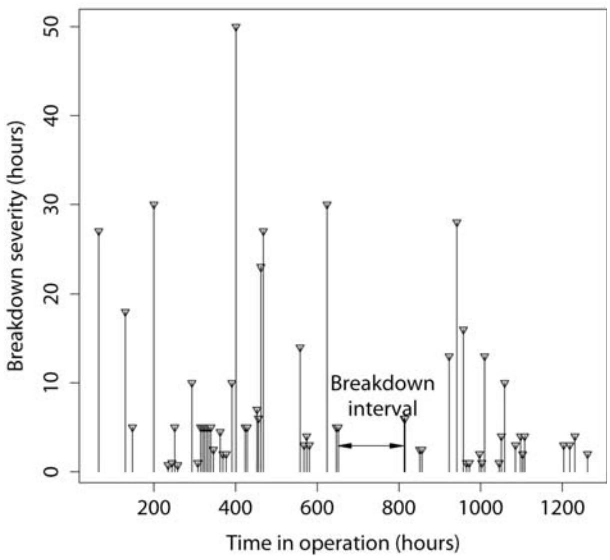

Figure 5: Excavation Interruptions Occurring between September 14, 2006 and May 10, 2007

The COSYE architecture for the simulation of the tunnel boring operations is comprised of several federates. One federate simulates the operations at the face of the tunnel, which include the excavation and the installation of the liners; another federate simulates the creation of tunnel sections; a third one handles the motion of the trains and the crane operations, etc. (see Ourdev et al. 2007) for details). The unexpected interruptions due to equipment failures and breakdowns are included in the breakdown federate, which implements the model described below. Figure 6 shows a plot of the breakdown characteristics, tracking the severity of the breakdown by the number of hours needed for repair and the time interval between breakdowns. The top histogram represents the marginal distribution of the severity, and the histogram on the right represents the marginal of the time interval.

4.2. Copula interruptions simulation

As pointed out in Section 2, the inference function for margin method involves two steps: first, fitting the marginals to the data, and then estimating the copula conditionally on the fitted marginals. The first step for finding the best fit for the marginals is to calculate the starting point for the numerical procedure. The data consist of $n = 58$ observations. The sample mean of the breakdown severity is $\overline{x}_S = 8.8$ hours and its variance is $\sigma_S^2 = 93.84$ hours. Substitution of these values into Equation 4 yields the following starting estimates for the parameters of the marginal distribution of the severity of the breakdowns: shape parameter, $\hat{\lambda}_{S0} = 2.419$, and rate parameter $\hat{\alpha}_{S0} = .299$. Similarly, we calculate the sample mean of the breakdown interval is $\overline{x}_F = 21.76$ hours and its variance is $\sigma_F^2 = 800.85$ hours. The substitution of these values into Equation 4 yields the following starting estimates for the parameters of the marginal distribution of the interval between the breakdowns: shape parameter, $\hat{\lambda}_{F0} = 1.510$, and rate parameter $\hat{\alpha}_{F0} = 0.069$.

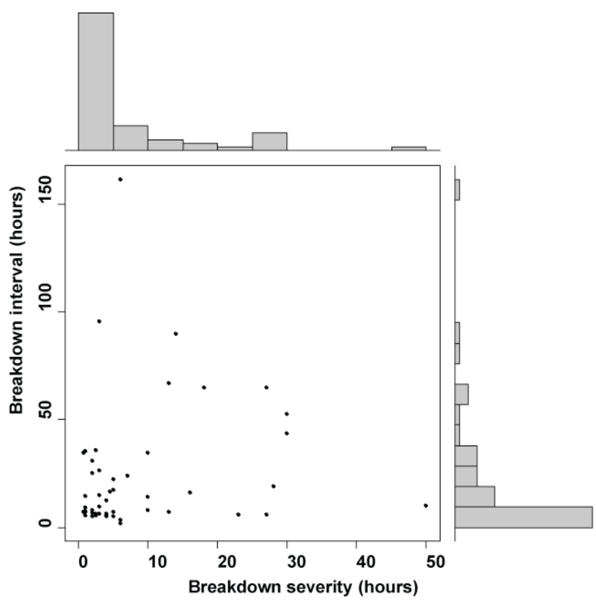

Figure 6: Plot of Breakdown Characteristics

Using the above initial estimates as starting points for the MLE procedure, we find that the best fit for the severity is given by a gamma distribution with parameters $S \sim Gamma(0.987, 0.153)$. The standard errors of the estimated parameters are: $SE(\hat{\alpha}_S) = 0.177$, and $SE(\hat{\lambda}_S) = 0.028$. For the frequency of the breakdowns we find another Gamma distribution with parameters $F \sim Gamma(1.1, 0.056)$. The standard error of the estimated parameters of this distribution are: $SE(\hat{\alpha}_F) = 0.177$, and $SE(\hat{\lambda}_F) = 0.010$. Figure 7 shows the histograms of the breakdown severity, and the intervals between breakdowns, with the corresponding probability densities of the best fits. Similarly, Figure 8 shows the empirical cumulative distribution functions (ecdf) of the breakdown severity, and the intervals between breakdowns, with the corresponding cumulative distributions of the best fits.

We used the Kolmogorov-Smirnov (KS) test to ascertain formally the goodness of fit of the above distributions. KS test quantifies the difference between the ecdf and the theoretical cdf, as shown in Figure 8 to formulate a hypothesis testing. The H_0 hypothesis is that the data comes from the specified distribution, versus the alternative, H_a, that the data is not from that distribution. The test statistics are formulated as the greatest difference between the ecdf, $F_{emp}(x)$, and the hypothesized theoretical cdf, $F_h(x)$

$$t = \sup_X \left| F_{emp}(x) - F_h(x) \right| \tag{21}$$

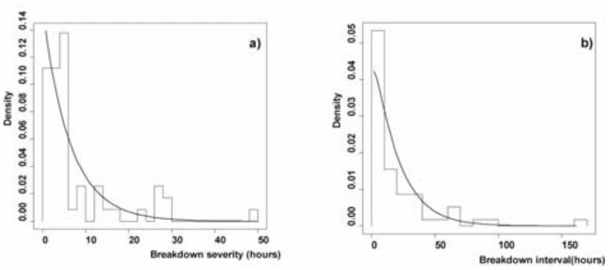

Figure 7: Histograms of the Breakdown Severity, Panel (a), and the Intervals Between Breakdowns, Panel (b), with the Corresponding Probability Density of the Best Fits

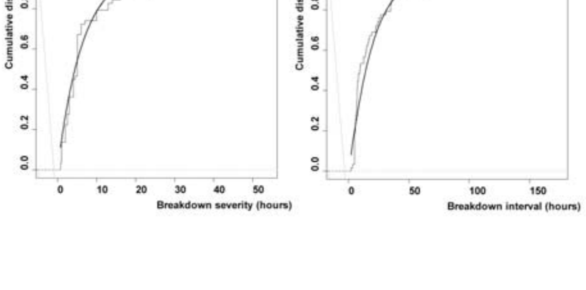

Figure 8: Empirical Cumulative Distributions of the Breakdown Severity, Panel (a), and the Intervals Between Breakdowns, panel (b), with the Corresponding Cumulative Distribution of Best Fits

The null hypothesis is rejected if the test statistics is greater than some critical value, or, alternatively, if the p-value is below the significance level.

The calculation for the KS test yields a p-value of $p_S = 0.268$ for the severity of the breakdowns, and a p-value of $p_F = 0.071$ for their frequency, so for both cases we cannot reject the null hypothesis at the significance level of $\alpha = 5\%$.

Having the best fits for the marginals, next, we proceed to find the best fit copula. Based on the observed two-dimensional data distributions, Figure 6, and the general considerations outlined above, we searched for a copula from the Student t copula class. We use the estimates of the shape parameters for the severity and frequency of the interruptions, $\hat{\alpha}_S$, and $\hat{\alpha}_F$, and the corresponding rate parameters, $\hat{\lambda}_S$, and $\hat{\lambda}_F$, obtained as described above as starting points of the maximum likelihood method.

The result of the MLE procedure brings a slight modification for the parameter of the marginals. The gamma distribution parameters for the severity become $S \sim Gamma(1.078, 0.132)$ with standard errors of $SE(\hat{\alpha}_S) = 0.176$, and $SE(\hat{\lambda}_S) = 0.028$. For the gamma distribution parameters for the frequency we find $F \sim Gamma(1.075, 0.048)$ with standard errors of $SE(\hat{\alpha}_F) = 0.176$, and $SE(\hat{\lambda}_F) = 0.010$. The correlation is estimated as $\rho = 0.267$ with a standard error of $SE(\hat{\rho}) = 0.135$. The contour plot for the resulting multivariate distribution is presented in Figure 9.

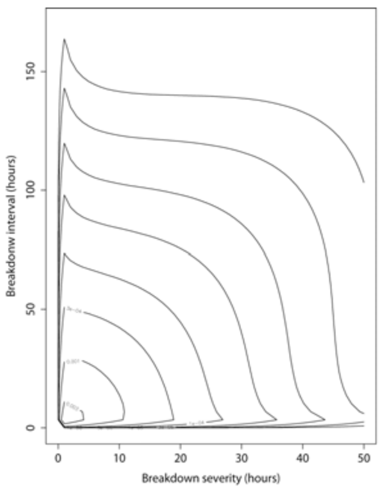

Figure 9: Contour Plot for the Multivariate Distribution Function Obtained as a Best Fit

With these estimated parameters, we can draw random samples from the interruptions distribution, to be used for the Monte Carlo simulation. The dots on Figure 10 represent 350 simulated values for process interruptions with severity and frequency obtained from the best fit to the observed interruptions as described above. The triangles on the figure visualize the actual observed to that moment interruptions and serve as additional check for the goodness of fit.
The procedure outlined above, involving the steps of fitting to the data and the Monte Carlo simulation, can be repeated every time a new interruption is registered by the system. Thus, the quality of the statistical fit will improve with the increase of the available data point. Such an online algorithm also allows the model to adapt to changes in the environment, such as equipment wear or variation among excavated strata.

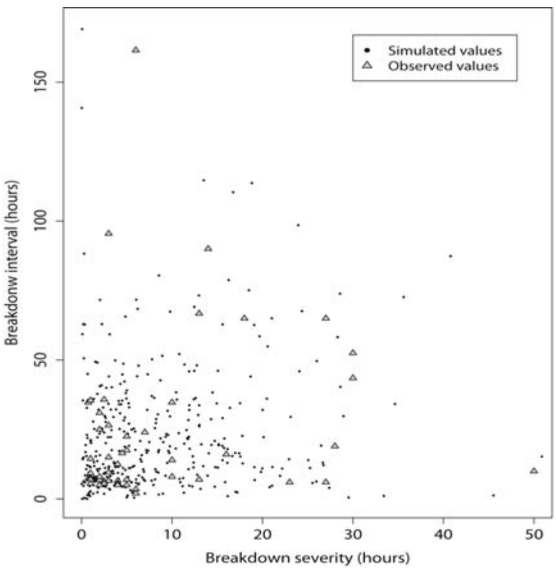

Figure 10: Simulated Values for Process Interrruptions with Severity and Frequency Obtained from the Best Fit to the Observed Interruption

5. CONCLUSIONS

In this paper we presented the first, to our knowledge, application of copulas to a construction project. We modeled the process interruptions that occur during a tunnel excavation by applying a system approach. In order to retain enough data for a meaningful statistical inference, we considered the operations comprising the excavation process as a system with the interruptions as one of the characteristic variables.

Although copulas are inherently multivariate objects, we restricted ourselves to the two-dimensional case. This allowed us to preserve the interpretation of one of the marginal distributions as the severity of the interruptions and the other as the frequency of interruptions (respectively: the interval between the interruptions). We carefully fitted the marginal distributions and the copula to the available data. The resulting two-dimensional probability distribution was used to generate random samples for a Monte Carlo simulation. The simulation results were used to estimate the changes in the project schedule and for more accurate and adaptive project management.

Our results showed the power of the copula approach for modeling and simulating uncertain dependent variables. The ease with which the copulas fit into the framework of Monte Carlo indicate a much broader application area, which would include more adequate risk modeling and risk management that does not rely on the assumption of normality.

APPENDIX A

The purpose of this appendix is to serve as an easy reference for comparison between some properties of the two-dimensional probability distributions and the copulas. The probabilistic approach to modeling uncertain quantities is to treat them as *random variables* (r. v.). Random variables, generally speaking, consist of two parts: the expected (most often occurring) value of the variable, and a measure of how uncertain we are about this expected value. The natural representation of a r. v., X, discrete or continuous, is its *cumulative distribution function* (cdf), defined as:

$$F_X(x) = P(X \leq x), \forall x \in \mathbb{R} \tag{A1}$$

For continuous random variable there is an alternative presentation, the *probability density function* (pdf), defined as the first derivative of cdf, i.e.

$$f_X(x) = F'_X(x) \tag{A2}$$

For a given pdf, $f_X(x)$ `$f_X(x)$`, the fundamental theorem of calculus allows calculating the corresponding cdf:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{A3}$$

It is very useful to introduce also the *quantile* function, which is the generalized inverse of the cdf, defined as:

$$F_i^{-1}(u) = \inf \left\{ x \mid F_i^{-1}(u) \geq x \right\}, \quad \forall u \in (0,1) \tag{A4}$$

For strictly increasing F_i, the quantile function F_i^{-1} becomes the ordinary inverse.

For a *pair* r. v., (X, Y), the dependence is completely described by their *joint cdf* defined as:

$$F_{X,Y}(x,y) = P(X \le x, Y \le y), \quad \forall (x,y) \in \mathbb{R}^2 \qquad \text{(A5)}$$

The equivalent probability model for the r. v. X and Y is given by their *joint pdf*, $f_{X,Y}(x,y)$ defined as:

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,du\,dv \qquad \text{(A6)}$$

The relation between the joint cdf and the joint pdf, corresponding to Equation 12 is given by:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x\,\partial y} \qquad \text{(A7)}$$

Integration over one of the r. v. yields the *marginal distribution* of the other, e.g.

$$f_X(x) = \int_{-\infty}^{y} f_{X,Y}(u,v)\,dv \qquad \text{(A8)}$$

Two random variables, *X* and *Y*, are *independent* if and only if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad \forall (x,y) \in \mathbb{R}^2 \qquad \text{(A9)}$$

The same proposition also holds in terms of cdfs, i.e.

$$F_{X,Y}(x,y) = F_X(x) F_Y(y), \quad \forall (x,y) \in \mathbb{R}^2 \qquad \text{(A10)}$$

REFERENCES

AbouRizk, S. M. and Mohamed, Y. 2000. *Simphony*: an integrated environment for construction simulation. *Proc. 2000 Winter Simulation Conf.*, IEEE, Orlando, Fla., 1907-1914.

Fishman, G. 2003. *Monte Carlo*, 4th ed. Springer, Cambridge, Mass.

Frees, W. E. and Valdez, E. A. 1998. Understanding relationship using copulas. *N. Amer. Actuarial J.*, 2, 1-15.

Genest, C. and Rivest, L.-P. 1993. Statistical interference procedures for bivariate archimedean copulas. *J. Amer. Stat. Assoc.*, 88423, 1034-1043.

Joe, H. and Xu, J. J. 1996 The estimation method of inference functions for margins for multivariate models. *Technical Report 166.* University of British Columbia, Vancouver, B.C.

Klugman, S. A., Panjer, H. H., Willmot, G. E. 2004. *Loss Models: From Data to Decisions*, 2nd ed. Wiley-Interscience, Toronto, Ont.

Malevergne, Y. and Sornette, D. 2005. Notions of copulas. *Extreme Financial Risks: From Dependence to Risk Management.* Springer, Cambridge, Mass.

McLeish, D. L. 2005. *Monte Carlo Simulation and Finance*. Wiley, Toronto, Ont.

Mikosch, T. 2006. Copulas: Tales and facts. *Extremes*, 9, 3-20.

Nelsen, R. B. 2003. Properties and applications of copulas: a brief survey. *Proc. 1st Braz. Conf. on Stat. Modeling in Insurance and Finance*, University of São Paulo, São Paulo, Brazil, 10-28.

Nelsen, R. B. 2006. *An Introduction to Copulas*, 2nd ed., Springer, Cambridge, Mass.

Ourdev, I., AbouRizk, S. M., and Al-Bataineh, M. 2007. Simulation and uncertainty modeling of project schedules estimates. *Proc. 2007 Winter Simulation Conf.*, IEEE, Washington, D. C.

SISC Simulation Interoperability Standards Committee of the IEEE Computer Society. 2000. *IEEE Standard for Modeling and Simulation M & S High Level Architecture HLA: Framework and Rules. Std. 1516-2000.*

Sklar, A. 1959. Fonctions de repartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*, 8, 229-231.

Wang, W. and Wells, M. 2000. Model selection and semiparametric inference for bivariate failure-time data. *J. Amer. Stat. Assoc.*, 95449, 62-72.

Yan, J. 2006. Multivariate modeling with copulas and engineering applications. *Handbook of Engineering Statistics*, H. Pham, ed., Springer, Cambridge, Mass., 973-990.