Proceedings of the 2007 Winter Simulation Conference
S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.

SIMULATION AND UNCERTAINTY MODELING OF PROJECT SCHEDULES ESTIMATES

Ivan Ourdev
Simaan Abourizk
Mohammed Al-Bataineh

Hole School of Construction, Department of Civil and Environmental Engineering,
5-080 Markin/CNRL Natural Resources Engineering Facility, University of Alberta,
Edmonton, Alberta, CANADA

ABSTRACT

Project management involves various sources of uncertainty that affect planning, execution schedules, and cost. At the same time, the influx of information can be employed to reduce the uncertainty. This can be efficiently accomplished within the framework of the Bayesian approach. This approach also has the advantage of providing a seamless synthesis of information coming from the field with information generated by data enhancing simulations. We demonstrate the use of this approach in an on-line simulation that augments a real-life monitoring and planning system for managing tunneling construction projects.

1 INTRODUCTION

Uncertainty is an inherent part of project management. It is critical for managing large high-risk projects, but can also directly affect the bottom line of relatively routine projects. In construction, engineering uncertainty and the concomitant risk lurk everywhere: uncertain durations, uncertain cost, sudden weather changes, equipment breakdown, human resource problems, unexpected changes in project scope, etc.

The most common casualty of uncertainty is the project *schedule*. Changes in the durations of specific tasks have a ripple effect on the start times of all consecutive tasks down the activity chain. Although a certain amount of *contingency time* is normally built in to the schedule of all projects, changes in the schedule have to be managed in a timely fashion in order to ensure the relatively smooth flow of labor and materials. This makes the forecast of task execution time an essential ingredient of successful project management.

An important general observation about evolutionary processes, and construction projects, in particular, is that the level of overall uncertainty normally decreases as time advances. This is due both to the decrease of the remaining project length and to the increase of the amount of available information about the project. As a result, an approach that suitably adapts the project variables to the arrival of new information could be very helpful in the adequate management of uncertainty. In this paper we employ the Bayesian method (see for example Gelman (2004), Lancaster (2004)) as an on-line tool for data analysis and forecast.

The Bayesian approach is a branch of the theory of random processes where the uncertain process quantities are considered *random variables* (r.v.) characterized by *probability density functions* (pdf). The approach uses the probabilistic framework to make statistical inferences about the ensemble averages of the random variables. An essential characteristic of Bayesian inference is the consistent method for updating the expected value of the r.v. in view of new evidence. Such updating can be done sequentially as the new evidence arrives and is very useful in building adaptive on-line monitoring and control systems.

In this paper we present an application of the Bayesian approach to a system for monitoring the productivity in a tunneling project and the forecasting of the progress in said project, called Construction Synthetic Environment (CoSyE). CoSyE is a discrete-event simulation system that gives the project planner the ability to produce effective project schedules and cost estimates. It allows for the simulation of all production operations with varying degrees of detail as well as modeling uncertain quantities as random draws from specified distributions.

The key element in the project is the *tunnel boring machine* (TBM) which drills tunnels of circular cross section. The production efficiency of the TBM is characterized by its *penetration rate*, defined as the distance drilled per unit of time. Knowledge of the historical penetration rate allows for the forecasting of the future position of the TBM and the ability to estimate its effect on the project schedule. The forecasting accuracy, admittedly, involves a high degree of uncertainty, being affected by changes in the soil type, variations of the water content, the degree of wear on the cutting edge, etc. Incorporation of new information as it arrives using the Bayesian method decreases

1-4244-1306-0/07/$25.00 ©2007 IEEE

uncertainty and provides a foundation for better management of project schedules.

This paper is organized as follows: In the next section we review the Bayesian methodology and describe the application of the Monte Carlo method for sampling. In section 3 we give an overview of the TBM operations that generate the data, while sections 0 and 5 describe how the data is modeled and the encompassing simulation framework. The last section presents a brief discussion of the possible ways of enriching the model.

2 UNCERTAINTY MODELING

2.1 Bayesian approach

The Bayesian approach has a long history of successful applications in enormously diverse disciplines (see for example Congdon (2007). The whole theory is built upon a single universally accepted mathematical proposition, Bayes' theorem, which asserts that the conditional probability that event A occurs given that the event B has already occurred is given by

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \qquad (1)$$

For a pair of random variables X and Y with marginal probability densities $p(x)$, and $p(y)$ and conditional densities $p(x \mid y)$, and $p(y \mid x)$ the theorem (1) is written as

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}. \qquad (2)$$

Usually y is interpreted as the observed 'data' and often is written as y^{obs}, while x plays the role of the vector of the parameters of the model, and is denoted by θ. The normalization constant in the denominator in (2), the marginal distribution of the data, does not depend on θ and is, usually, ignored, which leads to the following form of the Bayes' theorem:

$$\pi(\theta \mid y^{obs}) \propto p(y^{obs} \mid \theta)\pi(\theta). \qquad (3)$$

This is a mathematical expression of the proportionality of the *posterior* distribution $\pi(\theta \mid y^{obs})$ to the product of *prior* $\pi(\theta)$ distribution and the *likelihood* $p(y^{obs} \mid \theta)$, i.e. $\mathit{Posterior\ pdf} \propto \mathit{Likelihood\ function} \times \mathit{Prior\ pdf}$. Written in this form, Bayes' theorem provides a recipe for statistical inference. Here the uncertainty about the unknown parameters θ before making the observations y^{obs} is captured by the prior distribution $\pi(\theta)$. The information contained in the observations is incorporated in the model by applying Bayes' theorem (3) and results in the modification of the parameter uncertainty, modeled by the posterior distribution $\pi(\theta \mid y^{obs})$.

The prediction of unknown observables in the Bayesian framework is in terms of marginal distributions of data

$$\pi(y) = \int p(y \mid \theta)\pi(\theta)d\theta. \qquad (4)$$

This is called *prior predictive distribution* because it does not involve previous observations of the r.v. and only takes into account the uncertainty about the values of the parameters θ and the conditional uncertainty about the data y when θ are known. If the observations of a time-ordered random variable Y up to the moment t are $y_t^{obs} = (y_0, y_1, \cdots, y_t)$, then the value of a future observation y_{t+1} can be found from the *posterior predictive distribution*:

$$\pi(y_{t+1} \mid y_t^{obs}) = \int p(y_{t+1} \mid y_t^{obs}, \theta)\pi(\theta \mid y_t^{obs})d\theta. \qquad (5)$$

Traditionally, the Bayesian relied on symbolically tractable integrals by using *conjugate priors*. By definition, a class of prior distributions is a natural conjugate to a class of likelihood functions if the result from their multiplication posterior is a distribution of the same class as the prior. Popular examples are the pairs Normal-Normal, Poisson-Gamma, and Normal-Gamma, among others. Although the catalog of conjugate distributions is quite large, often, real life data is best modeled by combinations of distributions that are not conjugated. Fortunately, the increased power of computers made viable the alternative solution of *numerical integration* by the Monte Carlo method.

2.2 MCMC method

Integration is a key mathematical operation in the Bayesian approach. It is used to obtain the normalization constant in (3), to calculate marginal distributions as in (4) and (5), and to find the expected values of quantities of interest as $E_p[g(X)] = \int g(x)p(x)dx$, where g is some function of the r.v. X, which has a known pdf p.

The general idea of the Monte Carlo approach is to draw samples $\{x^{(i)}\}_{i=1}^N$ of size N from a *target* distribution $p(x)$ and calculate the mean of the integrand over the sampled points, i.e. $\overline{g}_N = \frac{1}{N}\sum_{i=1}^N g(x^{(i)})$. For i.i.d. samples by the law of large numbers $\overline{g}_N \to E_p[g(X)]$ as $N \to \infty$.

Critical factors for the accuracy of the Monte Carlo approach are the quality of the random number generator, and the sampling algorithm of the target distribution p. The most popular algorithms are importance sampling, rejection sampling, inversion, and Markov Chain Monte Carlo (MCMC); see for example Andrieu (2003). The last algorithm is particularly powerful and has already been implemented in various statistical packages.

The MCMC strategy uses a Markov-chain stochastic process with a stationary distribution that converges toward the required target distribution. The generated samples $\{x^{(i)}\}$ are identically, but not independently, distributed. The draws are sequential and each one depends on the previous value drawn with a distribution $x^{(t)} \sim p(x \mid x^{(t-1)})$ for $t = 1, 2, \cdots$ determined by the transition kernel p. Thus, at each step of the simulation we possess an approximation of the target distribution which is better than the approximation at the previous step.

There are various ways of constructing a Markov chain whose stationary distribution is the required target distribution. The most popular method is through the use of the Metropolis-Hastings (MH) algorithm, which starts from some crude *starting distribution* and proceeds to drawing *candidate points* x^* from a *proposal distribution* $x^* \sim p(x^* \mid x^{(t-1)})$. The candidate point is then accepted with *acceptance probability*

$$\alpha(x^*, x^{(t-1)}) = \min\left\{1, \frac{p(x^*)p(x^{(t-1)} \mid x^*)}{p(x^{(t-1)})p(x^* \mid x^{(t-1)})}\right\},$$

and rejected otherwise, i.e. retains its last value $x^{(t-1)}$. See Gelman (2004) for more details.

For practical problems involving complicated distributions, the sampling algorithm of choice is the Gibbs sampler, which uses the full conditional distributions $p(x_j \mid x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n)$ at step j of iteration t. The Gibbs sampling is interpreted as a special case of Metropolis-Hastings with acceptance probability $\alpha(x^*, x^{(t-1)}) = 1$. This interpretation allows the embedding of MH steps in the Gibbs algorithm when dealing with non-standard distributions. Otherwise, when the full conditional distributions belong to some standard distribution class (Normal, Beta, etc.) the samples are drawn directly.

3 DATA

The data is a subset of the information collected from the excavation of the tunnelling project SW3 executed in the City of Edmonton, Alberta, Canada using a tunnel boring machine. The project involves about 3.5km of sanitary sewer tunnels. It started in February of 2006 and is expected to finish in December of 2007. The tunneling operations are constantly monitored and data about the production progress is collected and all interruptions are recorded.

The tunneling operation comprises a set of activities, each one associated with a specific characteristic time. The activities sets are partitioned in cycles corresponding to the completion of one segment of the tunnel. The segment length is fixed to one meter corresponding to the length of the concrete cement liners used to cover the tunnel walls. Each cycle starts with unloading the liner blocks from the train. Usually two trains are used during the tunnel excavation, traveling back and forth in opposite directions between the entrance shaft and the face of the tunnel.

The empty train is used to collect the dirt from the excavation. After an excavation the length of one segment, the train travels back to the shaft while the TBM starts installing 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 will hoist the dirt from the sump pocket to the surface, where it is stockpiled. After dumping the dirt, the crane lowers down the liner blocks for the next segment of the tunnel. This completes one cycle of tunnel operations.

There are two sources of information about the daily production of the TBM. One is a surveying system called TACS, which gives the total duration for the installation of one segment. The other is the set of daily reports of the measurement of the total daily production in meters, including project delays and interruptions. The latter also contains information about the number of work-shifts per day (normally one or two), and the length of the shifts (normally ten or eight hours). The information from the daily reports is essential for a more accurate estimation of the actual proportion of production times recorded by the TACS.

After the synthesis of the information from these two sources, the obviously erroneous records are marked as missing (NA). All records with time durations shorter than the mean support time, or longer than one day if there is no corresponding information in the daily report, are ignored. This is done algorithmically by the data cleaning module of the CoSyE.

The available data for the period between 2006-09-14 and 2007-03-05 was used for building and testing the model. There are 545 time records in the TACS database and 134 corresponding daily productivity reports. The records were divided in two: a training set with a length of 460, and a test set of the remaining 85 records. The density distribution for the time duration over the full period between 2006-09-14 and 2007-03-05 is shown in Figure 1.

4 SIMULATION FRAMEWORK

The CoSyE simulation environment is a .NET implementation of the HLA (High Level Architecture) standard (Kuhl et al., 1999). The HLA architecture supports creation of complex virtual environments, called *federations*, using distributed simulation technologies. It provides a standard for combining individual components (*federates*) of such environment built by different people and maintaining the interoperability between them.

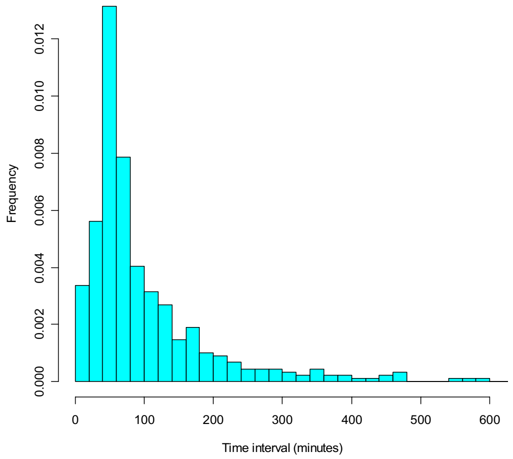

Figure 1 Density distribution of the time durations.

The CoSyE architecture is presented in Figure 2. Its core components are a Runtime Infrastructure (RTI) Server, an Object Model Template (OMT) editor, the system framework, and the modeling federates. The modeling federates can either be integral parts of the CoSyE system, or external software packages adding specific functionality.

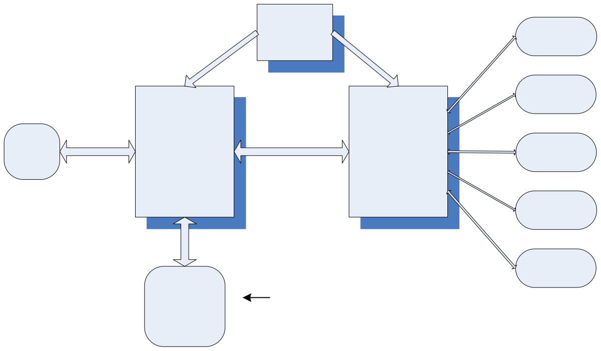

Figure 2 CoSyE architecture with the modeling federates comprising the tunnel boring simulation.

The simulation model of the tunnel boring operations is comprised of several federates. The *Excavation* federate simulates the operations at the face of the tunnel, which include both the excavation and the installation of the liners. The *Geotechnical* federate simulates the creation of tunnel sections using the data for the penetration rate. All operations involved in the removal of the excavated dirt from the tunnel, including the motion of the trains and the crane operations, are handled by the *Removal* federate. Equipment breakdowns are modeled as interruptions of the normal operation flow by the *Breakdown* federate. The *Statistical* federate collects relevant information from the model federates and produces summary reports, such as total duration to finish the tunnel, production per shift, equipment utilization, etc.

The foundation of the software architecture, the HLA, was designed specifically with the purpose of integrating diverse computer simulation systems. We employed this functionality to implement the penetration rate model using a separate simulation system, called WinBUGS (Spiegelhalter, 1996).

5 MODEL

The focus of the model is the uncertainty in the durations of the various production activities and their effect on the production rate of the TBM. From this point of view, all operations can be divided in two groups – production operations and supporting activities. The corresponding times spent in those operations are called production time and support time. The *production time* t_p is found as the difference between the total time needed to complete a section of the tunnel of length Δx (usually 1m), minus the *support time* t_s spent in supporting activities. Once the production time is known, the production rate is easily calculated as the $r_p = \Delta x / \Delta t_p$ in cm/min.

5.1 Support time

The support time has several components divided into two groups, depending on the degree of the uncertainty in their estimates. All support time is measured in minutes.

The first group consists of operations with relatively low variation in the estimation of the time it takes for completion. One such component is a constant that includes the time spent in shift start-up (15min) and shut-down (15min) as well as the 60min lunch time, in total $t_c = 90min$. Another component is the time it takes the train to travel the distance d between the entrance shaft and the current position of the TBM. It is calculated from the known average train velocity $V = 5km/h$, as $t_{tr} = d/V$ and increases linearly with time.

The second group is comprised of operations with a relatively high degree of uncertainty in their time duration. Their parameters are modeled as r.v. with empirical distributions determined on the basis of historical data collected during tunnelling and the experience of the personnel at the City of Edmonton (Ruwanpura *et al.,* 1999). Two of these components are modeled by the generalized *beta distribution* defined as follows:

$$Beta(x;a,b,\alpha,\beta) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta-1}}. \qquad (6)$$

For values of x in the interval between the location parameter, a, and the scale parameter, b, i.e. for $x \in (a,b)$, and for positive shape parameters, $\alpha > 0$, and $\beta > 0$. The beta function, $B(\alpha,\beta)$, is a part of the normalization constant and is typically expressed via the gamma function as $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.

One such high uncertainty component is the *lining time*, which is the time it takes to place the cement liners around the newly excavated section of the tunnel. It is modeled by a generalized beta distribution with parameters $t_{lin} \sim Beta(15,25,2,5)$ graphically presented in Figure 1(a). The time it takes to *load* the train is represented by symmetric generalized beta distribution $t_{load} \sim Beta(3,7,2,2)$, graphically presented in Figure 1(b). The time for unloading the train is approximately four times longer, i.e. it is $4t_{load}$, so the overall contribution of the loading and unloading operations to the total support time is $5t_{load}$. *The resetting time* is given by the uniform distribution $t_{res} = Unif(2,4)$ and presented for completeness in Figure 1(c).

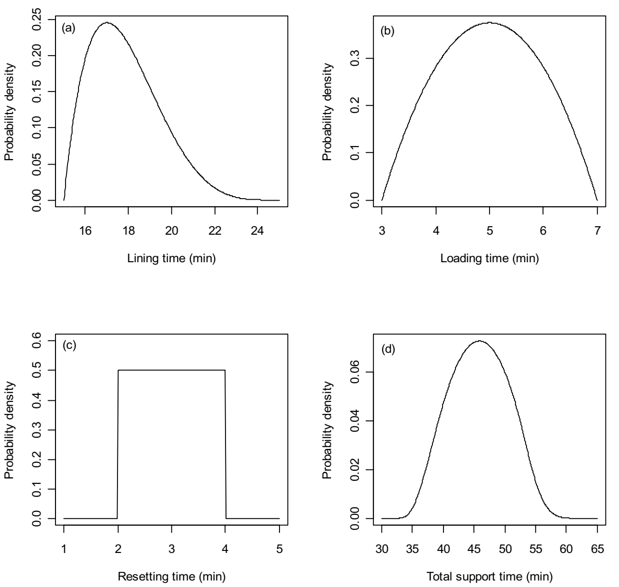

Figure 3 Density distributions of the various components of the support time; (a) lining time, (b) loading time, (c) resetting time, and (d) the variable component of the total support time.

The total support time t_s is the sum of all time intervals of the operations not directly involved in excavation

$$t_s = t_c + 2t_{tr} + t_{lin} + 5t_{load} + t_{res}\,. \tag{7}$$

Bearing in mind that the last three terms are independent random variables, this expression has to be interpreted as a convolution of the corresponding probability density functions. The resulting pdf for these three components of the total support time is graphically illustrated in Figure 1(d).

5.2 Penetration rate

The penetration rate r_t at time t was calculated for fixed distance increments $\Delta x = 1m$ as the ratio of the distance and production time. The latter was obtained by subtracting the support time (7) time from the observed total work-shift times recorded in the daily reports.

The data was modeled as an autoregressive process of the third order, AR(3), within the Bayesian approach as:

$$\mu_t = \beta_0 + \beta_1 r_{t-1} + \beta_2 r_{t-2} + \beta_3 r_{t-3} + \varepsilon_t\,, \tag{8}$$

with normally distributed penetration rates

$$r_t \sim Normal(\mu_t, \sigma^2)\,, \tag{9}$$

with mean μ_t, and variance σ^2. The order of the autoregressive process was suggested by the results from initial experiments with the models of different orders. The regression coefficients β_k were also assumed normally distributed

$$\beta_k \sim Normal(\mu_k, \sigma_k^2), \quad k = 0,\cdots,3\,, \tag{10}$$

with mean μ_k, and variance σ_k^2 fitted to the data.

The choice of this particular model was influenced by several factors. First, it was influenced by the need to incorporate and monitor the uncertainty of the inputs to the model. The second influencing factor was the requirement for adaptive updating of the model parameters. Thirdly, given the changing underground conditions and in particular the variation of the soil type, we wanted a model that on the one side reflects the historical values, but on the other, puts a higher weight on the more recent values. Autoregressive models of the type given by (8) adequately reflect the effect of the previous observations within the error margin ε_t. In addition, the Bayesian formulation allows the model parameters to be interpreted as random variables and the accuracy of the fit to be indirectly controlled.

The forecast of the average penetration rate for the next day was implemented as a two-step process. In the first step, all available data prior to the starting date was used to obtain the posterior distributions of the coefficients β_k of the autoregressive process, starting with non-informative priors:

$$\begin{aligned} &\beta_k \sim Normal(0,10^{-4}), \quad k = 0:3, \\ &\sigma^{-2} \sim Gamma(0.1,10^{-3}). \end{aligned} \tag{11}$$

Afterwards, the posterior predictive distribution (5) was found by sequential application of the Bayesian formula and informative priors for the parameters obtained from the previous iteration.

The mean values of the posterior coefficients of the model (8) along with the corresponding standard deviations and 95% confidence intervals (CI) are shown in Table 1.

Table 1: Posterior values for the autoregressive model and the corresponding standard deviations and 5% confidence intervals.

Node	Mean	Std. dev.	95% CI	
β_0	3.7820	0.5761	2.7100	4.9960
β_1	0.1392	0.0700	0.0036	0.2744
β_2	-0.0278	0.0742	-0.1878	0.1133
β_3	0.1155	0.0628	-0.0057	0.2364

Sequential application of the model to the out of sample data yields a standard error of 17%. Although the absolute value of the prediction error is significant, we consider this a promising result because of the high degree of uncertainty in the input values. Also, the large percentage of data records that were marked as missing or erroneous, 58% of the time durations in the test sample, after combining with the information from the daily reports must be noted.

6 DISCUSSION

The Bayesian method provides a powerful approach to decreasing uncertainty in project timelines and incorporating the impact of new information as it arrives. We applied this method for the on-line calculation of the penetration rate of a TBM within a distributed software framework that combines different sources of data from the field with discrete-event simulations. Still, the simulation can be improved in certain directions. For example, different operations have a different impact on the production rate. We expect that explicit separation of these effects would not only improve the accuracy of the forecast, but would also allow modeling the mean times of the duration of the effects, such as changes in the soil type, and the rate of wear on the cutting edge of the TBM. This would also allow for the implementation of a better algorithm for filtering out the outliers in the data from the field. Finally, we are planning on developing a stochastic model for the unplanned interruptions and breakdowns that also have a significant impact on the project timelines. The model will include simulation of both the mean time of failure and the mean time to repair.

REFERENCES

Andrieu, C., N. de Freitas, A. Doucet, and M. I. Jordan. 2003. An introduction to MCMC for machine learning. *Machine Learning*, 50: 5–43.

Congdon, P. 2007. *Bayesian statistical modeling*. 2nd ed. Wiley.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. Chapman and Hall/CRC.

Kuhl, F., R. Weatherly, and J. Dahmann. 1999. *Creating computer simulation systems: an introduction to the high level architecture*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Lancaster, T. 2004. *An Introduction to Modern Bayesian Econometrics*. Blackwell Publishing.

Ruwanpura J., S. AbouRizk, K. C. Er, and S. Fernando 1999. Simphony: Special Purpose Simulation Template for Utility Tunnel Construction. In *Proceedings of the 1999 Winter Simulation Conference*, Phoenix, AZ.

Spiegelhalter, D. J., A. Thomas, N. Best, and W. R. Gilks. 1996. BUGS: Bayesian inference using Gibbs sampling. Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge.

AUTHOR BIOGRAPHIES

IVAN OURDEV is a research associate in the Hole School of Construction at University of Alberta. His research interests lie in the area of uncertainty modeling and computer simulation of stochastic systems. His email address is <iourdev@ualberta.ca>

SIMAAN ABOURIZK is a Professor in the Department of Civil and Environmental Engineering at the University of Alberta. He holds the NSERC/Alberta Construction Industry Research Chair in Construction Engineering and Management, and the Canada Research Chair in Operational Simulation. He received his BSCE and MSCE in Civil Engineering from Georgia Institute of Technology in 1984 and 1985, respectively; and his Ph.D. degree from Purdue University in 1990. His research interests focus on the application of computer methods and simulation techniques to the management of construction projects. His email address is <abourizk@ualberta.ca> and his Web address is <http://www.construction.ualberta.ca/Faculty/abourizk.shtml>

MOHAMMED AL-BATAINEH is a Ph.D. candidate in the Department of Civil and Environmental Engineering at the University of Alberta. He received his BSc in Civil Engineering-Structural Engineering from Jordan University of Science and Technology, Jordan in 1999 and an M.Sc. in Construction Management from Western Michigan University in 2002. His research interests are in the application of simulation in construction management. His email address is <mta1@ualberta.ca>.