Bayesian Inference with Markov Chain Monte Carlo–Based Numerical Approach for Input Model Updating

Article *in* Journal of Computing in Civil Engineering · January 2020

DOI: 10.1061/(ASCE)CP.1943-5487.0000862

3 authors, including:

Lingzi Wu
University of Washington
26 PUBLICATIONS **164** CITATIONS
SEE PROFILE

Wenying Ji
George Mason University
69 PUBLICATIONS **767** CITATIONS
SEE PROFILE

Bayesian Inference with Markov Chain Monte Carlo–Based Numerical Approach for Input Model Updating

Lingzi Wu, S.M.ASCE[1]; Wenying Ji, A.M.ASCE[2]; and Simaan M. AbouRizk, M.ASCE[3]

Abstract: Stochastic, discrete-event simulation modeling has emerged as a useful tool for facilitating decision making in construction. Owing to the rigidity inherent to distribution-based inputs, current simulation models have difficulty incorporating new data in real-time, and fusing these data with subjective judgments. Accordingly, application of this valuable technique is often limited to project planning stages. To expand implementation of simulation-based decision-support systems to the execution phase, this research proposes the use of Bayesian inference with Markov chain Monte Carlo (MCMC)–based numerical approximation approach as a universal input model updating methodology of stochastic simulation models for any given univariate continuous probability distribution. Found capable of (1) fusing actual performance with expert judgment, (2) integrating actual performance with historical data, and (3) processing raw data by absorbing uncertainties and randomness, the proposed method will considerably improve the resilience, reliability, accuracy, and practicality of stochastic simulation models, thereby enabling the application of stochastic simulation in the execution phase of construction. **DOI: 10.1061/(ASCE)CP.1943-5487.0000862.** © *2019 American Society of Civil Engineers.*

Author keywords: Simulation-based analytics; Bayesian inference; Markov chain Monte Carlo; Input modeling; Real-time data analytics; Numerical method.

Introduction

As a tool, modeling has been widely used in engineering disciplines to design, analyze, communicate, test, and commission industrial, commercial, and residential facilities (AbouRizk et al. 2016). Simulation models are becoming increasingly used to support critical decision making in construction engineering. Of the myriad of simulation techniques available (Akhavian and Behzadan 2013), discrete-event simulation (DES) is most often applied in industrial and infrastructure construction decision-making processes owing to its ability to simulate resource interactions and operation logistics, especially for large and complex construction projects.

The success of a simulation model is highly dependent on accurately modeling the inputs, particularly in construction where a considerable number of inputs (each imbued with a wide variety of uncertainties) all relate to the underlying random process of various activities and tasks. The more accurate the model of the random input process, the more closely the simulation model mimics real-life behavior. To account for input variability, researchers have advocated for the modeling of inputs as probability distributions in a process known as stochastic or Monte Carlo simulation. Because of their ability to incorporate the randomness and various uncertainties inherent to construction activities, stochastic simulation models have been widely studied and used in the construction industry to enhance simulation-based decision-support systems.

Despite such advancements, the application of stochastic, discrete-event simulation models has traditionally been limited to the planning phase of construction. The industry continues to face notable challenges when it comes to adopting, upgrading, and using simulation models for decision support during the execution stage because inputs (e.g., a given distribution from historical data or experts' judgments) are often rigid, with no reliable or effective solution for fusing actual performance with the original input distribution to achieve real-time updating of the simulation model (Akhavian and Behzadan 2013). Because of these challenges, current simulation models have difficulty (1) reflecting real-time performance because of the use of static probability distributions; and (2) fusing subjective judgments with objective observations, thus limiting the application of simulation-based decision-support systems during the execution phase of a project. Although updating techniques, such as Bayesian statistics, have been proposed as a means of achieving real-time updating, many Bayesian-based methods require input data to have an analytical solution (i.e., conjugacy), limiting the application of these techniques in practice.

This study aims to address the limitation of real-time updating through the coupling of Bayesian inference with a Markov chain Monte Carlo (MCMC)–based numerical approximation approach, resulting in a universal input model updating method applicable to any univariate continuous probability distribution regardless of the conjugacy (i.e., a known parametric form of the posterior distribution). Demonstrated through its application on an illustrative case study, the proposed method was found capable of (1) fusing actual performance with expert judgment, (2) integrating actual performance with historical data, and (3) processing raw data by absorbing uncertainties and randomness. By enabling efficient, dynamic updating of the rigid inputs of a simulation model with new observations or subjective expert knowledge, the proposed method is

[1]Ph.D. Student, Dept. of Civil and Environmental Engineering, Univ. of Alberta, 9105 116 St., 5-080 NREF, Edmonton, AB, Canada T6G 2W2. Email: lingzi1@ualberta.ca

[2]Assisstant Professor, Dept. of Civil, Environmental and Infrastructure Engineering, George Mason Univ., 1411 Nguyen Engineering Bldg., 4400 University Dr., MS 6C1, Fairfax, VA 22030. Email: wji2@gmu.edu

[3]Professor, Dept. of Civil and Environmental Engineering, Univ. of Alberta, 9105 116 St., 5-080 NREF, Edmonton, AB, Canada T6G 2W2 (corresponding author). ORCID: https://orcid.org/0000-0002-4788-9121. Email: abourizk@ualberta.ca

Note. This manuscript was submitted on February 1, 2019; approved on April 19, 2019; published online on September 28, 2019. Discussion period open until February 28, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Computing in Civil Engineering***, © ASCE, ISSN 0887-3801.

expected to considerably improve the resilience, reliability, accuracy, and practicality of stochastic simulation models during the execution phase of construction.

Literature Review

Generalized Beta Family of Distributions

Following AbouRizk and Halpin's (1992) empirical study, which demonstrated the criticalness of using a flexible distribution (e.g., generalized beta distribution) to ensure the accuracy of the input modeling, the beta distribution has been extensively used for modeling inputs of the construction process over the last two decades. Among all of the flexible distributions, the generalized beta distribution with four parameters is one of the most widely recognized distributions for modeling construction processes (Chau 1995). Many researchers have successfully employed beta distributions to model a large number of construction management parameters including, but not limited to, the following: activity durations (Lu and AbouRizk 2000; Lu 2003; Poshdar et al. 2018; Zayed and Halpin 2001), construction costs (Inyim et al. 2016; Sonmez 2005; Wang et al. 2002), and quality management indicators (Ji and AbouRizk 2017).

Owing to its extensive usage in construction simulation modeling, the generalized beta family of distributions is presented here and is implemented in the case study. However, it is important to note that the proposed method is not limited to the beta distribution, and can be generalized to any other parametric probability distribution functions—a key contribution of the proposed method.

Mathematically, on an interval of $[L, U]$, a generalized beta distribution can be described as follows (AbouRizk et al. 1991; Ahsanullah 2017; Johnson et al. 1994):

$$f(y; a, b, L, U) = \frac{1}{B(a,b)} \cdot \frac{(y-L)^{a-1}(U-y)^{b-1}}{(U-L)^{a+b-1}}, \quad \text{if } L \le Y \le U$$

$$f(y; a, b, L, U) = 0, \quad \text{otherwise} \tag{1}$$

where $B(a, b)$ = beta function. With the transformation matrix shown as follows, $f(y; a, b, L, U)$ can be standardized to $f(x{:}a, b)$ with an interval of [0, 1]:

$$X = \frac{Y-L}{U-L}, \quad \text{if } L \le Y \le U \tag{2}$$

Standardized beta distribution is as follows:

$$f(x{:}a, b) = \frac{1}{B(a,b)} \cdot x^{a-1}(1-x)^{b-1}, \quad \text{if } 0 \le X \le 1$$

$$f(x; a, b) = 0, \quad \text{otherwise} \tag{3}$$

Thus, the generalized beta distribution can be treated as a standardized beta distribution with shape parameters $\{a, b\}$ scaled to the $[L, U]$ interval.

Bayesian Inference

The simulation models of construction processes developed in the previously mentioned studies have modeled their inputs based on either historical data or expert knowledge with fixed parameters as inputs and rigid assumptions. Construction processes, however, are highly dependent on the specific conditions that exist at the time they are performed, rendering them prone to deviation from expected baselines (Martinez 2010): what may have been anticipated and modeled in the planning stage of construction is often not what occurs during execution. The application of many construction simulation models proposed in the literature is, consequently, limited to the planning stages of construction.

Background

The Bayesian inference approach, developed by Bayes and Price (1763), has gained popularity in the twenty-first century due to its ability to incorporate multiple levels of randomness, integrate data originating from different sources, and reallocate credibility across the probability distribution of the value as new observations become available. Many researchers have since studied and demonstrated the practicality and benefits of implementing Bayesian techniques for updating underlying research interests (Brandley et al. 2015; Chung et al. 2004; Ji and AbouRizk 2017; Milo et al. 2015). While the aforementioned research was limited to conjugate priors, or a specific probability distribution, it has clearly demonstrated that a Bayesian approach can improve model accuracy, credibility, and reliability by systematically updating information of interest.

In contrast to Bayesian statistics, frequentist statistics suggests that the sampling process is "random," assuming that (1) the probability of each individual in the population being included in the sample is the same; and (2) separate drawings are mutually independent (Neyman 1937). It is generally agreed that "all scientific data has some degree of 'noise' in their value" (Kruschke 2014). Indeed, the underlying random processes in construction are associated with various uncertainties and conditions; however, achieving the *pure* randomness suggested by frequentists in applied statistics is impossible. Techniques used for data analysis should therefore be capable of inferring the underlying trends despite noise.

Bayesian statistics tackles the same problem from a different perspective. It systematically updates information of interest as more observations become available. Consequently, Bayesian inference is both flexible and practical owing to its ability to incorporate multiple levels of randomness and to combine information from various sources while absorbing all reasonable uncertainties in the inferential summaries (Gelman et al. 2013). Derived from Bayes's theorem, the basic components of Bayesian inference include the likelihood function, prior distribution, and joint posterior distribution. If prior distribution(s) are denoted as $p(\Theta)$ for the parameter set $\Theta = \{\theta_1, \ldots, \theta_n\}$, the likelihood function wherein all variables are related in a full probability mode denoted as $p(y|\Theta)$ and given a set of the new observation(s) of our underlying interest, $y = \{y_1, \ldots, y_n\}$, then the joint posterior distribution $p(\Theta|y)$ follows the numerical relation defined by Bayes's rule:

$$p(\Theta|y) = \frac{P(\Theta)p(y|\Theta)}{p(y)} \tag{4}$$

where $p(y) = \sum_{\Theta} p(\Theta)p(y|\Theta)$ for all possible values of Θ; or $p(y) = \int p(\Theta)p(y|\Theta)d\Theta$ for continuous Θ. Factor $p(y)$ is often called the marginal distribution of y or, more informatively, the prior predictive distribution (Gelman et al. 2013). Because it does not depend on Θ, and with fixed observation set y, it is a constant. Accordingly, the posterior distribution is proportional to the prior distribution multiplied by the likelihood function, denoted as

$$p(\Theta|y) \propto p(\Theta)p(y|\Theta) \tag{5}$$

Likelihood Function

In nonstatistical parlance, one could interchange *likelihood* for *probability*. Within Bayesian data analysis, however, probability

provides us the ability to predict unobserved data; likelihood, on the other hand, contains the available information through observed data (Statisticat 2013). Thus, the likelihood function is

$$p(y|\Theta) = \prod_{i=1}^{n} p(y_i|\Theta) \tag{6}$$

As a result, Bayesian inference obeys the likelihood principle, which states that the same likelihood function $p(y|\Theta)$ yields the same inference for parameter(s)-Θ for a given set of observations.

Prior Distributions

In Bayesian inference, a prior probability distribution (often referred to simply as a *prior*) of a parameter is a distribution that expresses uncertainty about the parameter before new observations are considered (Statisticat 2013). By applying Bayes's rule, the posterior distribution is affected by the selection of the prior distribution through multiplication. Consequently, the proper selection of the prior probability distribution strongly affects the outcome of the posterior distribution. Commonly, prior distributions are categorized into informative priors and uninformative priors, although further categorization has been suggested (Statisticat 2013). Where uninformative priors express minimal, vague, diffuse beliefs about the parameters, informative priors express specific information. If a project management team believes a current project is similar to a previous project, for example, priors that are similar to the historical data of a similar project could be defined. The model could thus consider both the historical data and current project performance.

Posterior Predictive Distributions

When making inferences about an unknown observation, the posterior predictive distribution is an indispensable component within the Bayesian data analysis of most practical problems. Given the observation data $y = \{y_1, \ldots, y_n\}$, to-be-observed data $\tilde{y}$ can be predicted using

$$p(\tilde{y}|y) = \int p(\tilde{y}|\Theta)p(\Theta|y)d\Theta \tag{7}$$

where $p(\tilde{y}|\Theta)$ = probability density function of $\tilde{y}$, given the fixed parameter(s)-Θ, which does not depend on observation y. To obtain the posterior predictive distribution $p(\tilde{y}|y)$, one must first sample parameter set Θ from the joint posterior distribution, and then simulate $\tilde{y}$ using $\tilde{y}^i \sim p(\tilde{y}|\Theta)$.

Application of Bayesian Inference for Real-Time Updating of Construction Models

Although Bayes's rule specifies the mathematical solution for the posterior distribution, exact analytical solutions rely on the possibility of computing the marginal probability. Historically, Bayesian inference techniques have been restricted to models with likelihood functions paired with corresponding formulas for prior distributions, known as conjugate priors (Kruschke 2014). Readers are referred to Jen and Hsiao (2018) for a detailed list of the most commonly used conjugate probability distribution functions.

Accordingly, it has been a longstanding challenge to generate simulation models that are capable of incorporating real-time updates during the execution phase (Akhavian and Behzadan 2013) to dynamically perform data-driven analytics and to provide critical decision-making support. Although research attempts have been made to use Bayesian techniques for real-time updating purposes in construction (Brandley et al. 2015; Chung et al. 2004; Ji and AbouRizk 2017; Milo et al. 2015), the methods and solutions proposed are limited to very specific cases. While Chung et al. (2004) proposed using a conjugate prior (i.e., normal distribution) for the probability density function (i.e., normal distribution) to achieve real-time updating of input models of a long-term repetitive tunneling project, the joint posterior distribution was assumed for the posterior predictive distribution. While both have the same mean in this case, the standard deviation differs; this results in an unrealistically small deviation. Later, Ji and AbouRizk (2017) carefully validated the Bayesian inference methodology on a binominal case for the quality control system of pipe fabrication. While a conjugate prior (i.e., beta distribution) for the Bernoulli likelihood function (i.e., binomial case) was presented, the study did not provide a universal solution for using the Bayesian inference methodology to update any univariate continuous input models regardless of conjugacy or likelihood function. Additionally, the need for fusing real-time performance with historical data to reflect the current project condition and integrating subjective expert opinion with objective observation remains unmet.

Performing Bayesian inference for realistic applications has often been limited to very specific cases, such as the aforementioned research, in which a prior distribution conjugate to the likelihood function is specified to yield an analytically solvable posterior distribution. However, with the development of random sampling algorithms (such as MCMC) and faster computer hardware, a broader selection of priors and likelihood functions are available for conducting Bayesian inference. With the help of MCMC and powerful computer hardware, an accurate approximation of the Bayesian posterior distribution is achievable in the absence of an exact analytical solution.

Markov Chain Monte Carlo

In cases where an analytical mathematical solution does not exist (i.e., where conjugacy cannot be met), a numerical approximation of the target distribution has been found to be a reliable alternative (Ji and AbouRizk 2017). The most commonly used approximation approach involves mimicking the target distribution through the random sampling of a large number of data points. Notably, in cases where the parameter space is relatively small, other approaches that systematically cover the parameter space by exhaustively computing the marginal probability can also be applied (Kruschke 2014).

Here, the MCMC method is used to generate an accurate approximation of the Bayesian posterior distribution, thereby providing a universal, real-time updating solution that overcomes previous limitations regarding conjugacy. The term *Markov chain Monte Carlo* combines two processes, namely, (1) Monte Carlo simulation, which involves the random sampling of a large number of values (Kroese et al. 2014); and (2) the Markov chain, "a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event" (Oxford Dictionaries 2019). The representativeness, accuracy, and efficiency of the MCMC method are attributable to both the algorithmic design of the method and the large number of iterations performed (Ji and AbouRizk 2017).

Of the many sampling algorithms, the Metropolis algorithm—developed in the 1950s by Metropolis et al. (1953) and further refined in the 1970s (Moller and Waagepetersen 2003)—has been widely used in physics, statistics, and applied sciences to approximate distributions (Robert and Casella 2011; Hitchcock 2003). Found capable of efficiently sampling single- and double-parameter problems, the Metropolis algorithm is well suited for sampling distributions commonly used to model construction processes and is, consequently, used here.

The steps of the Metropolis methods are demonstrated as follows:

1. Randomly generate a proposed leap, $\Delta\Theta \sim \text{normal}(\mu = 0, \sigma)$, and denote the proposed value of the parameter as $\Theta_{\text{proposed}} = \Theta_{\text{current}} + \Delta\Theta$.
2. Calculate the probability of moving to the proposed value

$$p_{\text{move}} = \min\left(1, \frac{p(\Theta_{\text{proposed}}|y)}{p(\Theta_{\text{current}}|y)}\right), \qquad p(\Theta|y) \propto p(\Theta)p(y|\Theta) \tag{8}$$

3. Accept the proposed parameter value if a random value sampled from a [0,1] uniform distribution is less than the p_{move}; otherwise, reject the proposed parameter value, and tally the current value again.

Methodology

This research proposes a method that couples Bayesian inference with an MCMC-based numerical approximation approach for updating univariate continuous probability distributions, regardless of the conjugacy. The proposed research method is illustrated in Fig. 1. To provide an illustrative example of the methodology, a generalized beta distribution with four parameters is outlined. It is important to note, however, that the proposed method can be applied to any univariate continuous probability distribution.

Step 1: Define the Probability Model

A descriptive probability model for all observable and unobservable quantities representing the underlying research interest (e.g., duration of a construction process, labor cost of activities, or productivity factor of a trade) is first defined. For illustrative purposes here, the underlying research interest is assumed to follow a generalized beta distribution $Y \sim \text{Beta}(y|a, b, L, U)$ with four parameters: a, b, L, and U.

Step 2: Identify and Understand the Parameters

The parameter(s) of the selected probability model (e.g., mean and standard deviation for a normal distribution, shape parameters for beta distribution) are then identified and understood. In the case of a generalized beta distribution, parameters L and U define the boundaries of the beta distribution. For example, if Y represents the duration of an activity, parameters L and U are the minimum and maximum durations of this activity recorded in historical data, respectively. In practice, the boundary parameters L and U are often well established, with no further updates required; hence, they can be considered constants. In contrast, shape parameters a and b, which directly control the shape of the beta distribution, commonly differ between projects; they are, therefore, the focus of the research interest.

Step 3: Specify Prior Distribution for the Parameters

In the case of multiple parameters, the credible values of the parameters may depend on the values of other the parameters, leading to a hierarchical model. Methods for addressing this issue are beyond the scope of this research. Alternatively, the credible values of the parameter may be independent of each other. For independent parameters a and b for a generalized beta distribution, the joint prior distribution follows: $p(a, b) = p(a)p(b)$. Based on the specific situation, informative priors (e.g., normal distribution) or uninformative priors (e.g., uniform distribution) can be chosen for $p(a)$ and $p(b)$.

Step 4: Bayesian Inference with MCMC Method

As more data points are collected, a Bayesian inference is conducted using the MCMC-based numerical method to derive the posterior distribution for the parameter(s). Given $Y \sim \text{Beta}(y|a, b, L, U)$, the probability of collecting new observation(s) $y_1, \ldots, y_n$ follows the mathematical form described as

$$p(y|a, b) = \frac{1}{B(a, b)} \cdot \frac{(y - L)^{a-1}(U - y)^{b-1}}{(U - L)^{a+b-1}}, \quad \text{if } L \le Y \le U \tag{9}$$

Considering a fixed set of observations, $y = \{y_1, \ldots, y_n\}$, $p(y|a, b)$ is the likelihood function of parameters a and b. Defined by Bayesian inference, the joint posterior distribution follows:

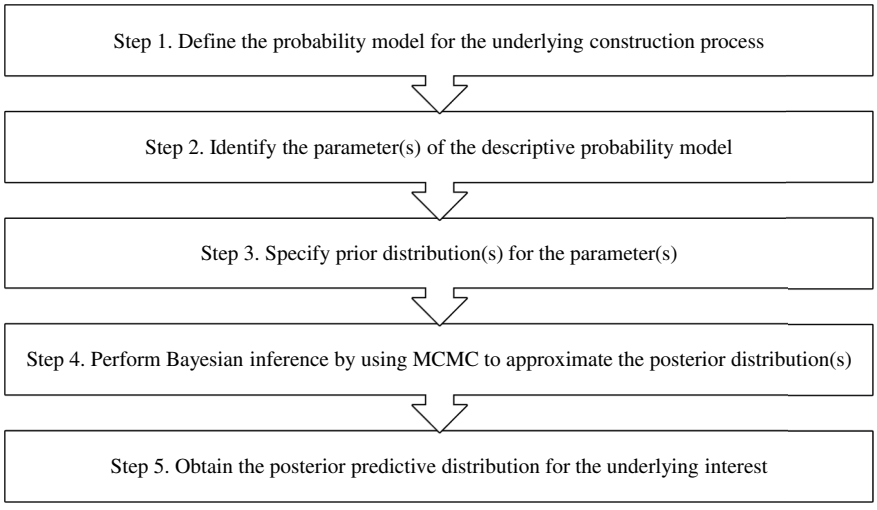

Fig. 1. Proposed methodology.

J. Comput. Civ. Eng., 2020, 34(1): 04019043

$$p(a,b|y) \propto p(a,b)p(y|a,b) = p(a)p(b)\prod_{i=1}^{n} p(y_i|a,b) \quad (10)$$

In both theory and practice, the log-likelihood is used instead of the likelihood on both the record level and model level. Thus

$$\log[p(a,b|y)] \propto \log[p(a,b)p(y|a,b)] = \log[p(a)] + \log[p(b)] + \sum_{i=1}^{n} \log[p(y_i|a,b)] \quad (11)$$

To approximate the joint posterior distribution, the MCMC numerical method with the Metropolis sampling algorithm will be applied as follows:

a. The approximation simulation begins with a set of initial values of parameters (a_1, b_1);
b. At the beginning of each iteration, randomly generate $\Delta a \sim \text{normal}(\mu = 0, \sigma_1)$ and $\Delta b \sim \text{normal}(\mu = 0, \sigma_2)$. Thus, $a_{\text{proposed}} = a_i + \Delta a, b_{\text{proposed}} = b_i + \Delta b$.
c. Calculate the probability of moving to the proposed value:

$$p_{\text{move}} = \min\left(1, \frac{p(a_{\text{proposed}}, b_{\text{proposed}}|y)}{p(a_i, b_i|y)}\right) = \min\left(1, \frac{p(a_{\text{proposed}})p(b_{\text{proposed}})\prod_{i=1}^{n} p(y_i|a_{\text{proposed}}, b_{\text{proposed}})}{p(a_i)p(b_i)\prod_{i=1}^{n} p(y_i|a_i, b_i)}\right) \quad (12)$$

d. Accept the proposed parameter values a_{proposed} and b_{proposed} if a random value sample from a [0,1] uniform distribution is less than p_{move}; otherwise, reject the proposed parameter values and return to Step 4b.

Following the completion of a desired number of iterations (e.g., 100,000), a set of samples for shape parameters a and b is generated. A histogram of the data set provides the reasonable representation of the joint posterior distribution $p(a,b|y)$.

Step 5: Obtain the Posterior Predictive Distribution

Finally, the posterior predictive distribution—representing the probability distribution of the yet-to-be-recorded data given the observed data—is derived. In the case of a beta distribution, the posterior predictive distribution for a future observation $\tilde{y}$ given y can be written as

$$p(\tilde{y}|y) = \iint p(\tilde{y}|a,b)p(a,b|y)dadb \quad (13)$$

To approximate the posterior predictive distribution, first sample a^i, b^i from the joint posterior distribution $p(a,b|y)$, then simulate $\tilde{y}^i \sim \text{beta}(a^i, b^i, L, U)$, where over time, $\tilde{y}_1, \ldots, \tilde{y}_n$ becomes an independently and identically distributed sample from $p(\tilde{y}|y)$ (Gelman et al. 2013).

Illustrative Case Study

Background

Because activity durations are one of the most studied and utilized inputs for simulation models of construction processes, a simplified simulation model of an earth-moving operation is used to demonstrate the feasibility and functionality of the proposed method. The simplified model captures a truck cycle that includes four major activities: loading, hauling, dumping, and return. The model simulates the delivery of 2,000 t of dirt using five 20-t capacity trucks that are loaded by shovels, which are assumed to be an unlimited resource, as illustrated in Fig. 2. Major activities and their durations are listed in Table 1. For the purposes of this case study, the duration of loading, dumping, and return are assumed to be constant, while hauling is assumed to follow a four-parameter generalized beta distribution fitted from experts' knowledge and historical observations.

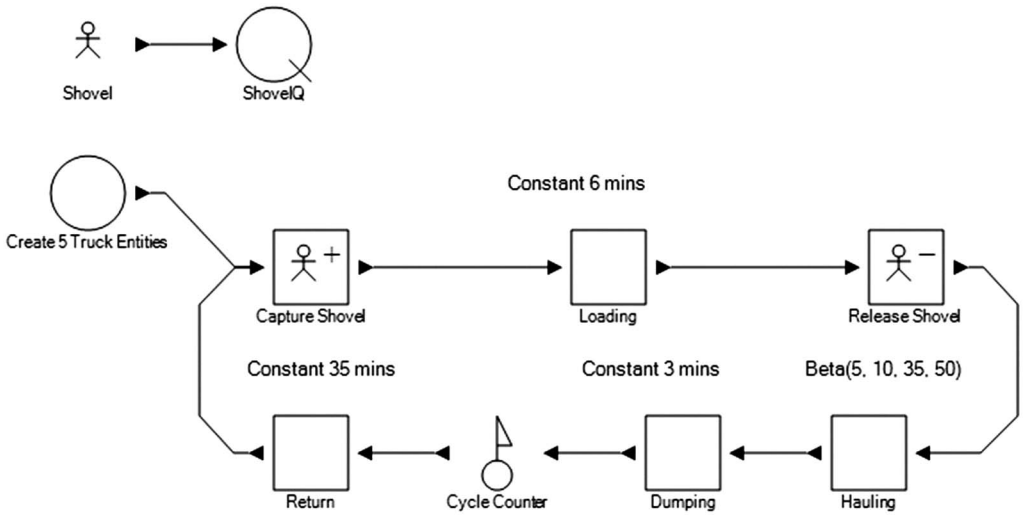

Fig. 2. Simulation model of simplified earth-moving operation.

Table 1. Original duration distributions of activities

Activity	Duration (min)
Loading 20-t truck	6
Truck hauling	Beta (5, 10, 35, 50)
Truck dumping	3
Truck return	35

Bayesian Updating of Input Models

While many construction processes are repetitive, it is uncommon to collect hundreds of observations between critical reporting and decision-making periods. Thus, an input updating method capable of generating reliable results from a limited number of data points is critical to be functional in practice. To demonstrate the ability of the proposed method to perform appropriately under such conditions, only 20 new observations were generated for each of the five reporting cycles. Overall, 100 new sample observations (Table S1) were randomly generated using the generalized beta distribution, Beta (5, 10, 35, 50). The proposed method was applied after 20 new observations were collected, and the accuracy of the proposed method was examined by comparing the input models derived using the proposed methodology (PM) with models that were directly fitted from the cumulative observations (CO), as well as the underlying distribution (UD).

Results

Shape parameters a and b that were obtained through direct fitting of the cumulative sampled observation actuals (i.e., in Cycle 1, 20 samples were used for fitting; in Cycle 2, 40 samples were used for fitting; etc.) are listed in the columns "Fitted on CO" and "Difference (% true) on CO" in Tables 2 and 3. Expectedly, the similarity of a and b to the underlying distribution increased as the number of data points accumulated. Notably, drastic fluctuations between cycles were observed, with the results of certain cycles being similar to the underlying distribution and others deviating considerably.

While the performance of this project is assumed to be similar to historical projects that follow the generalized beta distribution, Beta (5, 10, 35, 50), the project is characterized by certain unique features and uncertainties. Accordingly, a normal distribution, Normal (5, 0.5), was defined as the prior for shape parameter a, and a normal distribution, Normal (10, 1), as the prior for shape parameter b. This set of informative priors was chosen as a means of credibly considering historical data and expert opinion regarding uncertainty, where the mean value of each parameter was set at the most probable value based on the historical project data with the standard deviation representing 10% of the mean value to account for uncertainty. The posterior distribution of shape parameters a and b that was generated using the proposed method is listed in columns "Fit using PM" and "Difference (% true) using PM" in Tables 2 and 3.

Table 2. Shape parameters a fitted from CO versus PM

Cycle	True value	Fitted on CO	Difference (% true) on CO	Fit using PM	Difference (% true) using PM
1	5	8.7314	74.63	4.6963	6.07
2	5	5.1539	3.08	4.7936	4.13
3	5	4.3740	12.52	4.6823	6.35
4	5	4.0576	18.85	4.5827	8.35
5	5	4.2786	14.43	4.6661	6.68
Average		5.3191	24.70	4.6842	6.32

Table 3. Shape parameters b obtained using CO versus PM

Cycle	True value	Fitted on CO	Difference (% true) on CO	Fit using PM	Difference (% true) using PM
1	10	21.1659	111.66	10.3876	3.88
2	10	10.8684	8.68	9.9898	0.10
3	10	9.3608	6.39	9.9063	0.94
4	10	8.9133	10.87	9.9400	0.60
5	10	9.0413	9.59	9.8226	1.77
Average		11.8699	29.44	10.0093	1.46

The proposed method demonstrates considerable reliability between cycles, and accuracy when compared with the underlying distribution, especially given the small set of observations. The average percentage differences between the mean value of the posterior distribution and the true value for shape parameters a and b were 6.32% and 1.46%, compared with the direct fitting on the CO method, with 24.70% and 29.44%, respectively.

The histogram and trace plot of MCMC results for shape parameters a and b using PM in Cycle 1, together with the true values of the parameters, and parameters obtained through CO, are illustrated in Fig. 3. Histograms and trace plots for Cycles 2–5 are illustrated in Figs. S1–S4, respectively.

The results demonstrate that (1) the mean of the MCMC posterior samples for a and b (solid line) were more similar to the true parameter values (dashed line) when compared with the directly fitted from CO values (not represented in Fig. 3; represented in Figs. S1–S4) in all five cycles; and (2) the direct fitting from the CO method was associated with much larger fluctuations between cycles compared with the Bayesian inference (PM) method. Indeed, the parameter values fitted from CO were not within the presented scale for Fig. 3. In this instance, they are not shown.

The histograms of the samples of the posterior predictive distribution in Cycle 1, together with the three input model distributions, are illustrated in Fig. 4. Histograms for Cycles 2–5 are illustrated in Figs. S5–S8, respectively. Similar to the results of shape parameters a and b, the posterior predictive distribution was consistently closer to the underlying true distribution than the input distribution fitted directly from cumulative observations. The impact of the input modeling methods on project forecasting was also examined. The project was simulated using input models (1) directly fitted from the CO, (2) derived using the PM, or (3) using the UD. The forecasted project duration was determined for 1,500 runs; the results of the analysis are illustrated in Fig. 5.

Similar to the results obtained regarding the shape parameters and distributions, duration forecasts derived using the proposed input updating method were closer to the true underlying duration of the project for all five cycles when compared with the CO method. Moreover, during the first and second forecasting periods where the number of new observations was limited, the proposed method was found to more closely mimic the true underlying pattern and to more effectively incorporate various uncertainties (i.e., larger deviation window) than the direct fitting on the CO method. Indeed, the narrower deviation window of the CO method may result in an overoptimistic forecast, as observed in Fig. 5.

Sensitivity Analysis

To test the robustness of the proposed methodology, a sensitivity analysis designed to introduce a certain level of noise into the observation data to mimic the raw data collected from a real project

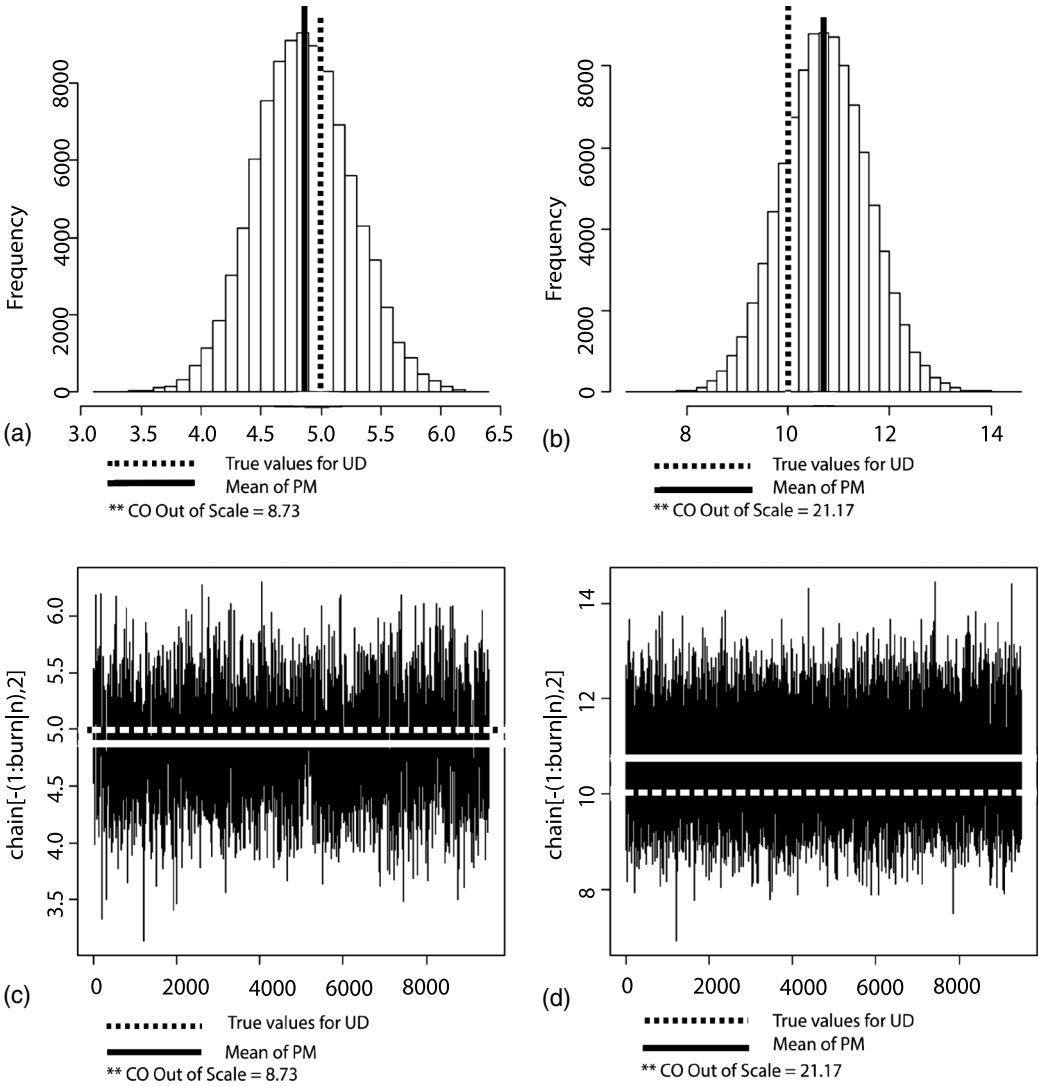

Fig. 3. (a and b) Posterior histogram; and (c and d) trace plot of parameters (a and c) *a* and (b and d) *b* for Cycle 1.

site was performed. One of the most common causes of fluctuations in productivity in construction projects is the learning effect, which is known to result in significant forecasting challenges in the early stages of a project. To mimic the noise of decreased productivity resulting from the learning effect, the simulated data points from the first three cycles (i.e., 60 random samples) of actual hauling duration were generated using 10% of the uniform distribution, Uniform (45, 50), and 90% of generalized beta distribution, Beta (5, 10, 35, 50), placing a higher probability of sampling a lower productivity. Assuming that after three cycles the project had achieved optimum productivity, the simulated data points for Cycles 4 and 5 (i.e., the remaining 40 random samples) were generated using the generalized beta distribution, Beta (5, 10, 35, 50). The 100 random samples that were generated using this approach are listed in Table S2. The shape parameters *a* and *b* fitted directly using cumulative observation samples are detailed in Tables 4 and 5 from columns "Fitted on CO" to "Difference (% true) on CO."

Similar to the base case study scenario, the similarity of *a* and *b* to the true values from the underlying distribution, Beta (5, 10, 35, 50), increased as the number of data points accumulated. With the introduction of noise, the direct fitting using CO method took longer to approach the underlying distribution, demonstrating that this approach is sensitive to the noise in the data set. While the differences in parameters *a* and *b* from the true values settled to around 10%–20% after a few cycles in the base scenario [column "Difference (% true) on CO" in Tables 2 and 3], the addition of noise resulted in a difference of around 20% for both parameters for all five cycles [column "Difference (% true) on CO" in Tables 4 and 5].

Taking into consideration the learning effect and the uncertainties associated with recorded data, the project was anticipated to follow the generalized beta distribution, Beta (5, 10, 35, 50). Again, informative priors were chosen with normal distributions, Normal (5, 0.25) and Normal (10, 0.5), as priors for shape parameters *a* and *b*, respectively. Because posterior distribution is influenced by both new observations and the prior distributions, a proper selection of the priors can affect the posterior given the same set of observations. If a set of uninformative priors is chosen, the posterior will show no influence from the priors but let the data speak for itself. To express firm belief in the subjective judgment of experts, the productivity fluctuation is caused by the learning effect, and the expected future observation will follow Beta (5, 10, 35, 50). The standard deviation was set as 5% of the value of the mean for both priors of parameters *a* and *b*. Corresponding posterior shape parameters are listed in Tables 4 and 5 from columns "Fit using PM" to "Difference (% true) using PM."

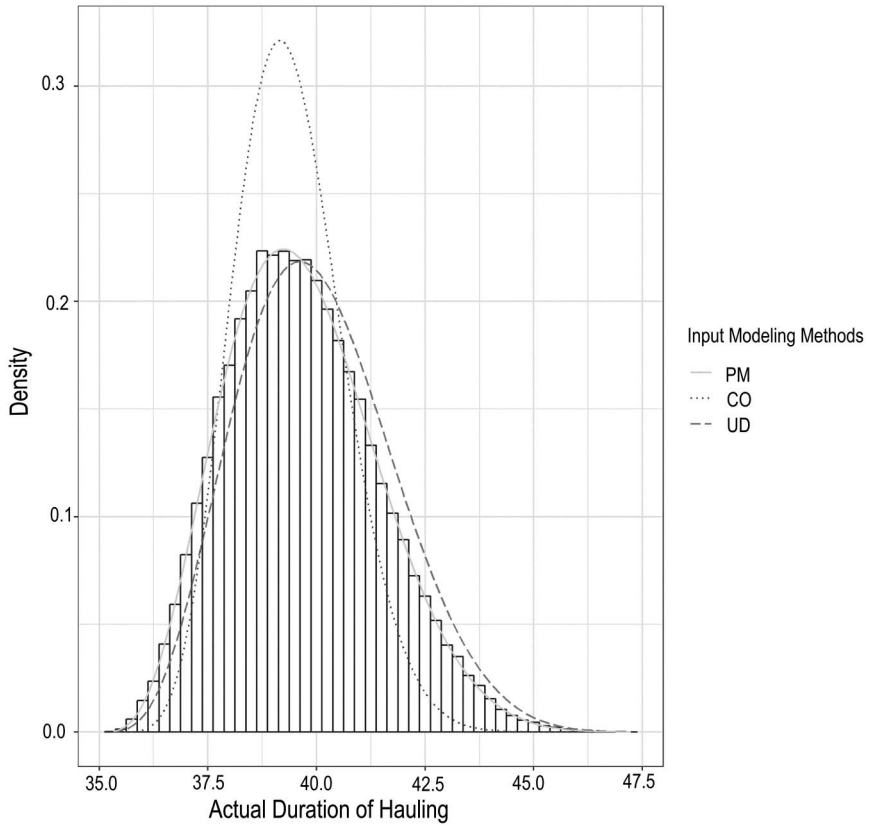

Fig. 4. Histogram of posterior predictive hauling model for Cycle 1.

As with the base scenario, the proposed method generated results that were more accurate and representative of the underlying probability distribution compared with the direct fitting using CO method. The average percentage differences between the mean value of posterior distribution and the true value for shape parameters a and b were 1.74% and 3.65%, compared with 24.96% and 23.21% fit directly from CO, respectively. The proposed method was also found to be comparatively insensitive to noise, with the average percentage difference similar for both the base scenario [column "Difference (% true) using PM" in Tables 2 and 3] and following the addition of noise [column "Difference (% true) using PM" in Tables 4 and 5]. To conclude, the proposed method

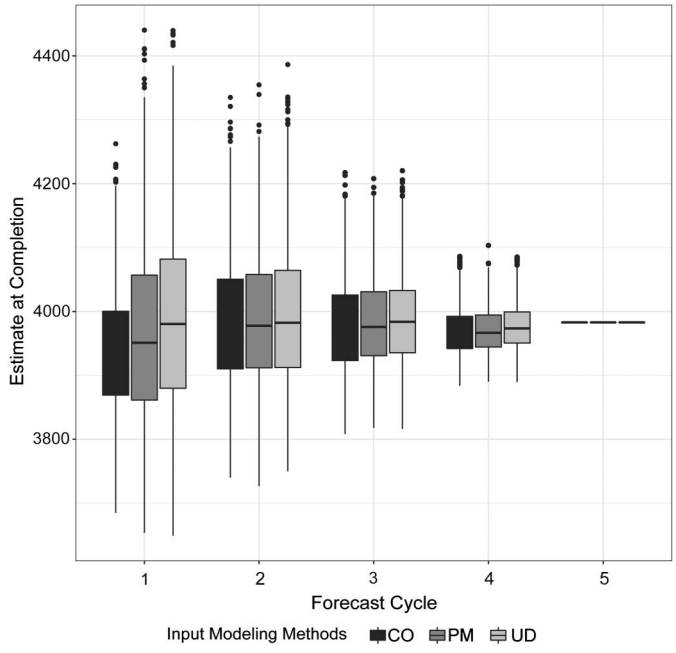

Fig. 5. Boxplot of simulation results obtained using input models directly fitted from the CO, derived using the PM, or derived using the UD.

Table 4. Shape parameters a fitted from CO versus PM

Cycle	True value	Fitted on CO	Difference (% true) on CO	Fit using PM	Difference (% true) using PM
1	5	7.0902	41.80	5.1683	3.37
2	5	4.2651	14.70	5.0146	0.29
3	5	3.4740	30.52	4.8798	2.40
4	5	3.7698	24.60	4.9439	1.12
5	5	4.3418	13.16	4.9255	1.49
Average		4.5882	24.96	4.9864	1.74

Table 5. Shape parameters b obtained using CO versus PM

Cycle	True value	Fitted on CO	Difference (% true) on CO	Fit using PM	Difference (% true) using PM
1	10	11.5305	15.31	9.7324	2.68
2	10	7.7220	22.78	9.6055	3.94
3	10	6.8282	31.72	9.7348	2.65
4	10	7.1323	28.68	9.5155	4.85
5	10	8.2430	17.57	9.5851	4.15
Average		8.2912	23.21	9.6347	3.65

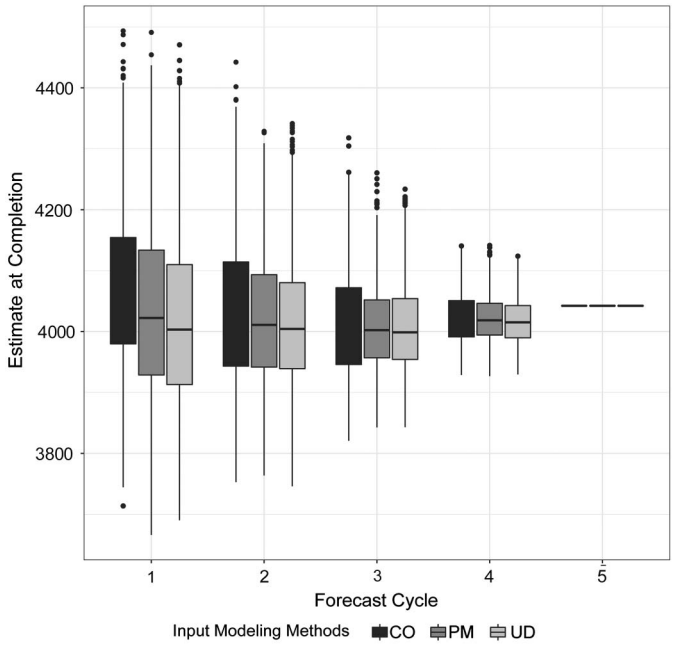

Fig. 6. Boxplot of simulation results obtained using input models directly fitted from the CO, derived using the PM, or derived using the UD.

demonstrated (1) robustness when the noise was introduced; and (2) desired representativeness and accuracy of both subjective opinion and objective observations.

The histogram and trace plot of MCMC results for shape parameters a and b using the proposed method, together with the true values of the parameters, and parameters obtained using the other aforementioned method, are illustrated in Figs. S9–S13. The histograms of the samples of the posterior predictive distribution, together with the three input model distributions, for Cycles 1–5, are illustrated in Figs. S14–S18. A comparison of the simulation results of project estimate at completion for each of the three input modeling methods is illustrated in Fig. 6.

As similarly, observed in the base scenario (Fig. 5), the simulation results obtained using the proposed input updating method were associated with less fluctuation in the presence of noise and generated more reliable duration forecasts compared with the direct fitting on cumulative observation method. This was particularly evident during Cycles 1 and 2, where the robustness of the proposed method and its ability to deal with limited and noisy data were most apparent.

Potential Applications

The implementation of the proposed methodology facilitates the dynamic real-time integration of data into the simulation models, thus enhancing the original model's accuracy and predictability. The traditional DES benefits from the real-time autocalibration of the input models by effectively assisting the decision-making process throughout both the project planning and execution phases of construction. This occurs in alignment with the dynamic data-driven application systems (DDDAS) philosophy (Darema 2004), which has also been referred to as simulation-based analytics (AbouRizk 2018; Ji and AbouRizk 2018b), and dynamic data-driven simulation (Ji and AbouRizk 2018a).

Potential realistic applications in construction engineering and management fields include but are not limited to production planning, earned value management, cost forecast, and risk management. Specifically, collected real-time performance data (e.g., production rate, productivity factor, actual cost) will be processed with the proposed methodology. The autocalibrated input models will then be utilized in simulation-based decision-support systems to reflect the dynamic project performance, deriving more accurate and meaningful decision-support output for practitioners. The proposed methodology can benefit any DDDAS, simulation-based analytics, or dynamic data-driven simulation developed for various engineering and applied science fields. For example, this method can be used to effectively process live sensor-generated data for real-time severe weather prediction, hazardous contamination production, traffic flow simulation, and so on.

Conclusions

Bayesian inference has been successfully implemented across many scientific and engineering disciplines to address the needs of multiple specific practical problems. However, many of the implementation methods, particularly in the area of construction engineering and management, are not generalizable because of their dependency on the availability of conjugate priors. Accordingly, many decision-support systems used in the construction industry remain unable to appropriately incorporate real-time information as it is generated.

This paper proposes a universal, Bayesian inference–based method for systematically updating any given univariate continuous probability distribution input model of simulations as new observations become available, and implements an MCMC-based numerical approximation approach to provide solutions regardless of conjugacy. An illustrative case study is used to demonstrate the generalizability, feasibility, and functionality of the proposed Bayesian inference with MCMC-based numerical method for updating simulation input models. The proposed method has been found capable of (1) effectively and efficiently updating input models as new observations become available; (2) accurately approximating the underlying probability distribution; (3) reliably fusing information from diverse sources, including subjective judgment and objective observations; (4) exhibiting robustness and resilience in situations where data were noisy and imbued with uncertainties; and (5) being generalized and applied to any given univariate continuous probability distribution. By applying the proposed method, input models of stochastic simulations can be effectively and efficiently updated in real time throughout the execution of a construction project.

The contributions of this research should be considered in light of several limitations. Owing to the nature of the illustrative case study, in which the random observations were generated based on a known underlying distribution, the fit of the model was not evaluated. In practice, where the underlying distribution is unknown, however, assessing the fit of the model to the data and to the subjective knowledge of experts after obtaining the posterior predictive distribution is essential (Gelman et al. 2013). Additionally, the selection of the prior distribution is a complex problem that requires consideration of historical data, professional experience, and regard for current project conditions. Proper prior distribution selection is of the utmost importance for ensuring the accuracy of the posterior distribution. Finally, while the proposed method provides a philosophical approach for integrating information from various sources, incorporating multiple levels of uncertainty and randomness, and consistently providing accurate, reliable results, the method itself does not represent a complete, decision-support system.

Laying the foundation for further dynamic, data-driven, simulation-based, and analytics-focused research in construction, future work building upon the proposed methodology is expected to result in a new generation of quantitatively driven, analytically based decision-support systems capable of providing real-time analytics, fusing various information sources, and incorporating randomness to enhance the efficiency and automation in construction management.

Acknowledgments

This research is funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Development Grant (CRDPJ 492657). The authors would like to thank Stephen Hague for sharing his knowledge and expertise of Bayesian inference.

Supplemental Data

Tables S1 and S2, and Figs. S1–S18 are available online in the ASCE Library (www.ascelibrary.org).

References

AbouRizk, S. M. 2018. "Simulation-based analytics: Advancing decision support in construction." In *Proc., 2nd European and Mediterranean Structural Engineering and Construction Conf.* Fargo, ND: ISEC Press.

AbouRizk, S. M., S. A. Hague, and R. Ekyalimpa. 2016. *Construction simulation: An introduction using Simphony.* Edmonton, Canada: Univ. of Alberta.

AbouRizk, S. M., and D. W. Halpin. 1992. "Statistical properties of construction duration data." *J. Constr. Eng. Manage.* 118 (3): 525–544. https://doi.org/10.1061/(ASCE)0733-9364(1992)118:3(525).

AbouRizk, S. M., D. W. Halpin, and J. R. Wilson. 1991. "Visual interactive fitting of beta distributions." *J. Constr. Eng. Manage.* 117 (4): 589–605. https://doi.org/10.1061/(ASCE)0733-9364(1991)117:4(589).

Ahsanullah, M. 2017. *Characterizations of univariate continuous distributions*. Paris: Atlantis Press.

Akhavian, R., and A. H. Behzadan. 2013. "Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining." *J. Constr. Eng. Manage.* 139 (11): 04013021. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000775.

Bayes, T., and R. Price. 1763. "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S." *Philos. Trans.* 53: 370–418. https://doi.org/10.1098/rstl.1763.0053.

Brandley, R. L., J. J. Bergman, J. S. Noble, and R. G. McGarvey. 2015. "Evaluating a Bayesian approach to demand forecasting with simulation." In *Proc., 2015 Winter Simulation Conf.*, 1868–1879. New York: IEEE.

Chau, K. W. 1995. "Monte Carlo simulation of construction costs using subjective data." *Constr. Manage. Econ.* 13 (5): 369–383. https://doi.org/10.1080/01446199500000042.

Chung, T. H., Y. Mohamed, and S. AbouRizk. 2004. "Simulation input updating using Bayesian techniques." In *Proc., 2004 Winter Simulation Conf.*, 1238–1243. New York: IEEE.

Darema, F. 2004. "Dynamic data driven applications systems: A new paradigm for application simulations and measurements." In *Proc., Int. Conf. on Computational Science*, 662–669. Berlin: Springer.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Hitchcock, D. B. 2003. "A history of the Metropolis–Hastings algorithm." *Am. Stat.* 57 (4): 254–257. https://doi.org/10.1198/0003130032413.

Inyim, P., Y. Zhu, and W. Orabi. 2016. "Analysis of time, cost, and environmental impact relationships at the building-material level." *J. Manage. Eng.* 32 (4): 04016005. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000430.

Jen, H., and C. Hsiao. 2018. "Using Bayesian inference modeling in estimating important production parameters used in the simulation-based production planning." In *Proc., IEEE Int. Conf. on Applied System Innovation 2018*, 1038–1041. New York: IEEE.

Ji, W., and S. M. AbouRizk. 2017. "Credible interval estimation for fraction nonconforming: Analytical and numerical solutions." *Autom. Constr.* 83 (Nov): 56–67. https://doi.org/10.1016/j.autcon.2017.07.003.

Ji, W., and S. M. AbouRizk. 2018a. "Data-driven simulation model for quality-induced rework cost estimation and control using absorbing Markov chains." *J. Constr. Eng. Manage.* 144 (8): 04018078. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001534.

Ji, W., and S. M. AbouRizk. 2018b. "Simulation-based analytics for quality control decision support: A pipe welding case study." *J. Comput. Civ. Eng.* 32 (3): 05018002. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000755.

Johnson, N. L., S. Kotz, and N. Balakrishnan. 1994. *Continuous univariate distributions*. 2nd ed. New York: Wiley.

Kroese, D. P., T. Brereton, T. Taimre, and Z. I. Botev. 2014. "Why the Monte Carlo method is so important today." *Wiley Interdiscip. Rev. Comput. Stat.* 6 (6): 386–392. https://doi.org/10.1002/wics.1314.

Kruschke, J. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London: Academic Press.

Lu, M. 2003. "Simplified discrete-event simulation approach for construction simulation." *J. Constr. Eng. Manage.* 129 (5): 537–546. https://doi.org/10.1061/(ASCE)0733-9364(2003)129:5(537).

Lu, M., and S. M. AbouRizk. 2000. "Simplified CPM/PERT simulation model." *J. Constr. Eng. Manage.* 126 (3): 219–226. https://doi.org/10.1061/(ASCE)0733-9364(2000)126:3(219).

Martinez, J. C. 2010. "Methodology for conducting discrete-event simulation studies in construction engineering and management." *J. Constr. Eng. Manage.* 136 (1): 3–16. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000087.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. "Equation of state calculations by fast computing machines." *J. Chem. Phys.* 21 (6): 1087–1092. https://doi.org/10.1063/1.1699114.

Milo, M. W., M. Roan, and B. Harris. 2015. "New statistical approach to automated quality control in manufacturing processes." *J. Manuf. Syst.* 36 (Jul): 159–167. https://doi.org/10.1016/j.jmsy.2015.06.001.

Moller, J., and R. P. Waagepetersen. 2003. *Statistical inference and simulation for spatial point processes*. Boca Raton, FL: CRC Press.

Neyman, J. 1937. "X—Outline of a theory of statistical estimation based on the classical theory of probability." *Philos. Trans. R. Soc. London, Ser. A* 236 (767): 333–380. https://doi.org/10.1098/rsta.1937.0005.

Oxford Dictionaries. 2019. "Definition of Markov chain in US English." Accessed January 21, 2019. https://en.oxforddictionaries.com/definition/us/markov_chain.

Poshdar, M., V. A. González, G. M. Raftery, F. Orozco, and G. G. Cabrera-Guerrero. 2018. "Multi-objective probabilistic-based method to determine optimum allocation of time buffer in construction schedules." *Autom. Constr.* 92 (Aug): 46–58. https://doi.org/10.1016/j.autcon.2018.03.025.

Robert, C., and G. Casella. 2011. "Short history of MCMC: Subjective recollections from incomplete data." *Stat. Sci.* 26 (1): 102–115. https://doi.org/10.1214/10-STS351.

Sonmez, R. 2005. "Review of conceptual cost modeling techniques." *AACE Int. Trans.* ES71–ES74.

Statisticat. 2013. "Bayesian inference." Accessed January 31, 2019. https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/BayesianInference.pdf.

Wang, L., W. Shen, H. Xie, J. Neelamkavil, and A. Pardasani. 2002. "Collaborative conceptual design—state of the art and future trends." *Comput.-Aided Des.* 34 (13): 981–996. https://doi.org/10.1016/S0010-4485(01)00157-9.

Zayed, T. M., and D. Halpin. 2001. "Simulation of concrete batch plant production." *J. Constr. Eng. Manage.* 127 (2): 132–141. https://doi.org/10.1061/(ASCE)0733-9364(2001)127:2(132).