Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

INTRODUCTION TO SIMULATION INPUT MODELING

Bahar Biller

Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213, U. S. A

Canan Gunes

Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213, U. S. A

ABSTRACT

In this tutorial we first review introductory techniques for simulation input modeling. We then identify situations in which the standard input models fail to adequately represent the available input data. In particular, we consider the cases where the input process may (i) have marginal characteristics that are not captured by standard distributions; (ii) exhibit dependence; and (iii) change over time. For case (i), we review flexible distribution systems, while we review two widely used multivariate input models for case (ii). Finally, we review nonhomogeneous Poisson processes for the last case. We focus our discussion around continuous random variables; however, when appropriate references are provided for discrete random variables. Detailed examples will be illustrated in the tutorial presentation.

1 INTRODUCTION

An important step in the design of a stochastic simulation is to represent the random inputs of the system being studied. Examples of random inputs include the time to failure for a machining process, the demand per unit time for inventory of a product, and the times between arrivals of calls to a call center. Input modeling is the practice of selecting probability distributions (i.e., input models) to represent such random input processes. It is important for the simulation analyst to recognize that there is no "true" input model for any stochastic input; the goal is to obtain an approximation that captures the key characteristics of the underlying input process.

Input modeling problems can be categorized into two broad classes depending on the presence or absence of data. When data are available, we fit a probability distribution to the data. Otherwise, all information available is to be used for constructing an input model. In this tutorial, we discuss both cases. When data is present, we assume the availability of continuous data, but we provide references for discrete data. Past WSC tutorials including Nelson and Yamnitsky (1998), Biller and Nelson (2002), Leemis (2004), and Kuhl et al. (2009) provide additional discussion to the material presented here. Biller and Nelson (2002) answer ten input modeling questions that are likely to be asked by the simulation practitioners. Leemis (2004)'s survey is an example-driven tutorial describing the basic steps of fitting a probability distribution to the service time data and the arrival time data, while Kuhl et al. (2009) particularly focus on methods for fitting flexible univariate input models to the data. Nelson and Yamnitsky (1998)'s tutorial is more advanced compared to others as it also discusses input modeling when there is correlation in the input process.

We organize the remainder of the paper as follows. In Section 2, we describe the basic steps of fitting a standard input model to the available data; we call the input models that are included in most commercial simulation software packages the standard input models. Also discussed are the construction of an empirical distribution for the available data and fitting an input model when no data

are available. In Section 3, we present input models that are used when standard input models are inadequate. In particular, Section 3.1 reviews flexible input models; Section 3.2 discusses multivariate input models, while Section 3.3 reviews input processes changing over time. We conclude in Section 4.

2 INPUT MODELING

Fitting with data is composed of three main steps: (1) Select one or more candidate probability distributions, based on physical characteristics of the process and graphical examination of the data. (2) Determine the values of the input model parameters. (3) Check the goodness of the fit via tests and graphical analysis. We review each of these steps in Section 2.1, Section 2.2, and Section 2.3, respectively. In Section 2.4, we describe how to construct an empirical input model when none of the standard input models is a good fit to the available data. Finally, we switch our attention to the case where no data are available in Section 2.5.

2.1 Picking an Input Model for the Data

Many commercial input modeling packages has thirty to forty standard probability distributions (e.g., Stat::Fit has 32 different distributions, BestFit has 37 probability distributions, while ExpertFit has 40 distributions) from which to choose an input model for the data. The reason that there are many choices is that each distribution is developed to represent certain physical processes. For example, normal distribution can be used for representing the time to assemble a product that is simply the sum of the times required for each (sub)assembly operation, while exponential distribution is well suited for capturing the time to failure for a system that has constant failure rate over time. Banks et al. (2001, Chapter 9) provide a description of the physical bases of various standard distributions.

The physical basis, if available, is a very valuable information in choosing an input model for the data; however, we might not have this information. In these cases, most software packages fit a number of distributions to the data depending on some input from the user such as whether the data is discrete or continuous, whether there are bounds on the range of the random variables. Then, they form a list of candidate distributions that could be used to represent the data, estimate the parameters of the candidate distributions, rank the distributions according to their goodness of fits, and finally recommend the top distribution in the list having the best goodness-of-fit measure. In Section 2.2, we describe the parameter estimation of the selected input model, while in Section 2.3 we discuss widely used goodness-of-fit tests for choosing among different input distributions.

2.2 Estimating the Parameters of the Input Model

After a probability distribution is chosen, the next step in input modeling is to estimate the parameters of the selected probability distribution. Common methods for parameter estimation are the maximum likelihood estimation method, the method of matching moments, the method of matching percentiles, and the least-squares estimation method (Law 2007). All simulation input modeling packages estimate the parameters of the fitted input model using either one or a couple of these methods. The method used does not generally matter as long as the software estimates the parameters with numerically stable algorithms. Good sources for parameter estimation are Banks et al. (2001, Chapter 9) and Law (2007, Chapter 6).

2.3 Assessing the Goodness of the Fit

Common methods for checking how well a fit captures the key characteristics of the available input data include the goodness-of-fit tests such as Kolmogorov-Smirnov and Anderson-Darling tests, and plots such as density-histogram plots and probability plots. Specifically, both the Kolmogorov-Smirnov and the Anderson-Darling test compare the fitted distribution to the empirical distribution. However, the Kolmogorov-Smirnov test is designed to emphasize discrepancies in the middle of the data, while the

Anderson-Darling test is designed to emphasize the discrepancies in the tails. The density-histogram plots compare the fitted density function to the histogram of the data, while probability plots compare the quantiles of the fit with the quantiles of the data.

The goodness-of-fit tests are built on hypothesis testing. While the null hypothesis states that the selected input model along with its parameter estimates is a good fit for the data, the alternative hypothesis claims the opposite. The test rejects the fit only if there is overwhelming evidence that the selected input model is not a good representative of the data. One measure used to summarize the result of the hypothesis testing is the $p-$value, which is the significance level at which one would reject the null hypothesis for the given data. A large $p-$value (> 0.1) supports the null hypothesis that the distribution is a good fit for the data on hand. In comparing among different input models that could be used to represent the available data, many software packages use the Kolmogorov-Smirnov and the Anderson-Darling tests. Let $F_n(x)$ represent the empirical distribution function for the available data of size n such that $F_n(X_{(i)}) = i/n$ for $i = 1,2,\ldots,n$, and $\hat{F}(x)$ is the fitted distribution function. The Kolmogorov-Smirnov test statistic D_n is the largest vertical distance between $F_n(x)$ and $\hat{F}(x)$ for all values of the available data; i.e., $D_n = \sup_x\{|F_n(x) - \hat{F}(x)|\}$, and it can be computed as $D_n = \max\{D_n^+, D_n^-\}$, where $D_n^+ = \max\{i/n - \hat{F}(X_{(i)})\}$ and $D_n^- = \max\{\hat{F}(X_{(i)}) - (i-1)/n\}$ for $i = 1,2,\ldots,n$. Alternatively, the Anderson-Darling test statistic A_n is a weighted average of $[F_n(x) - \hat{F}(x)]^2$, and is computed as $A_n^2 = (-\{\sum_{i=1}^n (2i-1)[\text{In}(\hat{F}(X_{(i)})) + \text{In}(1 - \hat{F}(X_{(n+1-i)}))]\}/n) - n$. Many input-modeling packages report D_n and A_n for several fitted candidate distributions; the lower D_n and A_n suggest smaller distance between the empirical distribution and the fitted distribution, and thus a better fit.

Despite their popularity, the goodness-of-fit tests should be used with caution since they are unlikely to reject any distribution when little data is available, and are likely to reject every distribution when lots of data are available. This makes sense because when a single data point is available, any distribution would be a reasonable fit, while when data are abundant, the test can easily identify the discrepancies between the data and the fitted distribution. Another shortcoming of the goodness-of-fit tests is that they represent the lack of fit by a single number, while plots specify where the lack of fit occurs. Therefore, the goodness-of-fit tests are not enough themselves; one also needs to check the fit via plots. A widely used plot is the density-histogram plot, which compares the histogram of the available data with the density of several fitted distributions. We let X_1, X_2, …, X_n represent the available data of size n, and break the range of the data into k disjoint intervals $[b_0, b_1)$, $[b_1, b_2)$, …, $[b_{k-1}, b_k)$ of equal width; i.e., $\triangle b = b_j - b_{j-1}$ for $j = 1,2,\ldots,k$. An histogram is obtained by plotting the sorted data versus the function $h(x)$ defined as

$$h(x) = \begin{cases} 0 & \text{if } x < b_0, \\ h_j & \text{if } b_{j-1} \le x < b_j \text{ for } j = 1,2,\ldots,k, \\ 0 & \text{if } x \ge b_k, \end{cases}$$

where h_j is the proportion of X_j values that fall in the j^{th} interval $[b_{j-1}, b_j)$. A density-histogram plot compares $h(x)$ with $\triangle b\hat{f}(x)$, where $\hat{f}(x)$ is the fitted probability density function. Despite being available in many input-modeling packages, density-histogram plots are sensitive to how we group the data and may lead to different conclusions depending on the width $\triangle b$ of the histogram cells. Another graphical comparison method is the probability plots. There are two well-known probability plots: the quantile-quantile ($Q-Q$) plot and the probability-probability ($P-P$) plot. The $Q-Q$ plot displays the sorted data $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ versus the fitted quantile $\hat{F}^{-1}((i-0.5)/n)$ for $i = 1,2,\ldots,n$, and a straight line implies that the fitted distribution is a good model for the available data. The $P-P$ plot, on the other hand, is obtained by plotting the fitted probability $\hat{F}(X_{(i)})$ versus the sample probability $\tilde{F}_n(X_{(i)})$, and if $\hat{F}(X_{(i)})$ and $\tilde{F}_n(X_{(i)})$ are close to each other the plot will be approximately linear. Thus, the $Q-Q$ plot amplifies the difference between the tails of the fitted model $\hat{F}(x)$ and the sample distribution function $\tilde{F}_n(x)$, while the $P-P$ plot amplifies the difference between the middle of the fitted model $\hat{F}(x)$ and the sample distribution function $\tilde{F}_n(x)$. We refer the

reader to Vincent (1998) and Law (2007) for an excellent discussion on checking the goodness of a fit.

2.4 When None of the Standard Distributions is a Good Fit

When no standard distribution provides a good fit, we use the empirical distribution of the data for representing the input uncertainty. The data may be available in two forms: the individual observations are available or only the number of observations that fall into pre-specified regions is available. We focus on the former case here where the individual observations are available, and refer the reader to Section 6.2.4 of Law (2007) for a discussion of the latter case. When the individual data points $X_1, X_2, \ldots, X_n$ are available, we can specify the empirical cumulative distribution function (cdf) $F(x)$ for this input data set as

$$F(x) = \begin{cases} 0 & \text{if } x < X_{(1)}, \\ \frac{i-1}{n-1} & \text{if } X_{(i)} \le x < X_{(i+1)} \text{ for } i = 1,2,\ldots,n-1, \\ 1 & \text{if } x \ge X_{(n)}, \end{cases}$$

where $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ represent the sorted data (Law 2007). As the number of data points n increases, the empirical cdf converges to the "true" cdf. One shortcoming of this method is that only discrete values of the sample can be generated for the execution of the simulation; i.e., this method does not produce values that are between the observed data points. Various interpolation methods can be used to fill the gaps and smooth the empirical cdf allowing sampling values between the observations (see Banks et al. (2001, Chapter 8) for one of them), but they often miss the tail behavior. One way to represent the tail behavior is to attach an exponential cdf to the right-hand side of the empirical cdf. The resulting cdf is called the mixed empirical-exponential cdf (Bratley et al. (1987)) and is given by

$$F(x) = \begin{cases} 0 & \text{if } x < X_{(1)}, \\ \frac{i}{n} + \frac{(x - X_{(i)})}{n(X_{(i+1)} - X_{(i)})} & \text{if } X_{(i)} \le x < X_{(i+1)} \text{ for } i = 1,2,\ldots,n-k, \\ 1 - \frac{k}{n}\exp(-(x - X_{(n-k)})/\theta) & \text{if } x \ge X_{(n-k)}. \end{cases}$$

The typical practice is to cut off from one to five $(1 \le k \le 5)$ observations to form the tail and choose θ to make sure that the mean of the distribution is $\bar{X} = \sum_{i=1}^{n} X_i/n$; i.e., $\theta = (X_{(n-k)}/2 + \sum_{j=n-k+1}^{n}(X_{(j)} - X_{(n-k)}))/k$. We refer the reader to Section 6.2.4 of Law (2007) for building an empirical distribution for discrete data.

2.5 When No Data are Available

When no data are available, the key idea is to use any available information that may help to identify some characteristics of the process. Examples of the available information include expert opinion, physical and conventional limits, and the nature of the process itself. Expert opinion is generally valuable because experts are the people who are familiar with the process being modeled, and thus they are likely to provide the center value and the extreme values for the process. Furthermore, if we know the upper and the lower bounds of the process (if any), any distribution chosen to model this process will respect to these bounds. Similarly, if we know that we are modeling the number of customers that arrive to a bank during a period of time with a constant arrival rate, then the nature of this process implies that the Poisson distribution would be a good model for this process.

A widely used approach in the absence of data is to use the expert opinion. Depending on the level of detail the expert provides, different probability distributions can be used to model the process. For instance, assuming that we are trying to model the demand process of a product in our inventory and we are given by the expert that demand is likely to be no less than 100 units but no more than 500 units, then we can use a uniform distribution with lower bound 100 and upper bound 500 to

model this demand process. If we are additionally provided by the expert that the demand is most likely to be 350 units, then we can use a triangular distribution with parameters 100, 350, and 500 to model this demand process. The ideal information that could be obtained from the expert would be the breakpoints in addition to the minimum and the maximum values of the process. Specifically, breakpoints are values and an each value is associated with a percentage of being less than itself. For instance, for the demand process, the expert may say that the demand of the product will be at least 100 units, it has a 50% chance of being more than 300 units, and a 15% chance of beating 450 units, while it will definitely not exceed 500 units. These breakpoints then can be used to construct a piecewise continuous distribution function that could be incorporated into simulation packages.

Instead of specifying some breakpoints, some experts may opt to specify a mean value and a percent variability around it. For instance, the expert may say that the demand of the product was 100 units; this year they are expecting a 15% increase with plus or minus 20%. We can use normal distribution with mean 115 and standard deviation 0.2×115 to model this demand process. However, one should be careful here in interpreting the statements of the experts. Although experts provide a mean value for the process, what they provide as mean may indeed be the most likely value of the process. Therefore, it is better to ask for both the mean value and the most likely value, and interpret the answer accordingly. Furthermore, the standard deviation (e.g., plus or minus 20%) provided by the expert may not reflect the average deviation from the mean but the most extreme deviation. Similarly, for precise modeling, one should ask the expert for both the average deviation and the most extreme deviation and make sure that we obtain the correct information from the expert.

3 INPUT MODELING WHEN STANDARD MODELS ARE INADEQUATE

Despite being widely used due to their availability in commercial software packages, limited shapes of the standard distributions may not be flexible enough to represent key characteristics of the data. System inputs may exhibit dependence either in time sequence and/or with respect to other, violating the independence assumption of the standard input models. Also, the input process may change over time. In Section 3.1, we review flexible input models that can be used for overcoming the first drawback of the standard input models. In Section 3.2, we consider the second drawback of the standard input models and review two widely used multivariate input models. Finally, in Section 3.3 we discuss the use of the nonhomogeneous Poisson process for representing input processes changing over time.

3.1 Flexible Input Models

To overcome the drawbacks of the standard distributions, many flexible distributions have been suggested in the literature including the curves proposed by Pearson (1895), the Johnson translation system (Johnson 1949a), the generalized lambda distribution (Ramberg and Schmeiser 1974), the four-parameter distribution introduced by Schmeiser and Deutsch (1977), the generalized beta family (Kuhl et al. 2009), and the Bézier distribution (Kuhl et al. 2009). The four-parameter Schmeiser-Deutsch distribution is particularly easy to use, but the one-to-one relationship between the distribution parameters and the moments is lost. They also fall short of capturing the distributional characteristics such as bimodality and heavy tails. The generalized beta family represents random variables that have bounded supports, while the generalized lambda family matches any first two moments but limited third and fourth moments. However, it is particularly easy to represent the key characteristics of data with the generalized beta family as the mean and variance of this family are simple functions of its parameters. The translation system developed by Johnson (1949a) has the ability to capture any finite first moments of a distribution while representing a wide variety of unimodal and bimodal distributional shapes. DeBrota et al. (1989) discuss the advantages of the Johnson translation system in comparison with triangular, beta, and normal families of distributions. The curves of the Johnson translation system and the Pearson's system generally agree with each other; however, the mathematical structure

of the distributions from the Johnson translation system provides a convenient aid to the estimation and generation of input processes for stochastic simulations.

Specifically, a random variable X from the Johnson translation system has a cdf of the form

$$F(x) = \Phi\left\{\gamma + \delta f\left[\frac{x-\xi}{\lambda}\right]\right\},$$

where γ and δ are shape parameters, ξ is the location parameter, λ is the scale parameter, $\Phi(\cdot)$ is the cdf of the standard normal variable with mean zero and variance one, and $f(\cdot)$ is one of the following transformations:

$$f(y) = \begin{cases} y & \text{for the } S_N \text{ (normal) family,} \\ \log(y) & \text{for the } S_L \text{ (lognormal) family,} \\ \log(y/(1-y)) & \text{for the } S_B \text{ (bounded) family,} \\ \log(y+\sqrt{y^2+1}) & \text{for the } S_U \text{ (unbounded) family} \end{cases}$$

Within each family, a distribution is completely specified by the values of the parameters ξ, λ, γ, and δ, and the range of X depends on the family of interest: $X > \xi$ and $\lambda = 1$ for the S_L family; $\xi < X < \xi + \lambda$ for the S_B family; and $-\infty < X < \infty$ for the S_N and S_U families. There is a unique family; i.e., choice of f, for each feasible combination of the coefficient of skewness and the coefficient of kurtosis that determines parameters γ and δ. We refer the reader to Johnson (1987) and DeBrota et al. (1989) for the distributional shapes that can be represented with the Johnson translation system.

Fitting data to Johnson translation system typically consists of two consecutive steps: (1) The selection of the appropriate family (S_N, S_L, S_B or S_U) that best represents the available input data. (2) The determination of the parameters of the selected Johnson family. Two widely used methods for performing both of these tasks are the method of matching moments (Hill et al. 1976) and the method of matching percentiles (Slifker and Shapiro 1980). Another fitting method that is particularly designed for estimating the parameters of the selected Johnson distribution is the least-squares estimation method of Swain et al. (1988). More specifically, the FITTR1 software developed by Swain et al. (1988) can be used for fitting the Johnson translation system via the least-squares estimation method as well as the method of matching moments, method of matching percentiles, and the minimum L_1 and L_∞ norm. The VISIFIT software of DeBrota et al. (1989a) fits Johnson's S_B family to the subjective information when no data are available.

Random variate generation via Johnson translation system is easily accomplished by transforming a standard normal variable Z into $X = \xi + \lambda f^{-1}[(Z-\gamma)/\delta]$, where $f^{-1}(a) = a$ for the S_N family, $f^{-1}(a) = \exp(a)$ for the S_L family, $f^{-1}(a) = 1/(1+\exp(-a))$ for the S_B family, and $f^{-1}(a) = (\exp(a) - \exp(-a))/2$ for the S_U family.

3.2 Multivariate Input Modeling

Almost all simulation input-modeling packages assume the availability of independent and identically distributed (i.i.d.) input data. However, input processes may exhibit dependence. Vincent (1998) provides techniques to check whether dependence exists in a data set. Dependence can occur in time sequence or across different input processes, or both. The input processes can be categorized into three groups depending on the form of dependence: (1) univariate time series, (2) vector times series, and (3) random vector.

A univariate time series represents a sequence of random variables that all have the same probability distribution, but exhibit serial dependence. This type of dependence is represented by the autocorrelation, which is the correlation between observations within the series. Examples of time series input models include the monthly demands of a product of a retailer, and the interarrival times of customers to a queue. The vector time series input model, on the other hand, represents dependence both in time sequence and across components. For instance, it represents the dependence between

both the monthly demands of a product and the demands of different products. We refer the reader to Biller and Ghosh (2006) for a discussion of time series input models and vector time series input models.

A random vector models each random input of a stochastic simulation by a different probability distribution, while allowing the inputs to be dependent on each other. This type of dependence is represented by a correlation matrix of pairwise correlations. Random vectors are often used as input models for the demands of different products of a retailer and the annual returns on different asset classes of portfolios. In this section, we review two commonly used random vector input models, the multivariate normal distribution (Section 3.2.1) and the multivariate Johnson translation system (Section 3.2.2). The multivariate normal distribution is a special random vector whose all components are normally distributed, while the multivariate Johnson translation system represents a random vector where each component can belong to one of the families (S_N, S_L, S_B or S_U) of the Johnson translation system. Thus, the multivariate Johnson translation system is a more flexible multivariate distribution than the multivariate normal distribution in representing the key characteristics of the data on hand. A review of multivariate input models including multivariate Bézier distribution families and multivariate generalized beta family of distributions can be found in Kuhl et al. (2010), while Biller and Gunes (2010) discuss multivariate Poisson distribution and the multivariate negative binomial distribution as the two widely used multivariate discrete distributions.

3.2.1 Multivariate Normal Distribution

A $k-$dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)'$ with mean vector μ and covariance matrix Σ is said to have a nonsingular multivariate normal distribution if Σ is positive definite and the joint probability density function of the normally distributed components X_i, $i = 1, 2, \ldots, k$ is of the form

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

(Tong 1990). The availability of a statistically valid fitting algorithm (Johnson and Wichern 2001) and a fast sampling algorithm (Press et al. 2007) often motivates the simulation practitioner to use the multivariate normal distribution for driving the stochastic simulation with correlated inputs. Despite the ease in its implementation, the main limitation to the use of multivariate normal distribution for simulation input modeling is that the marginal distribution of each component is normal and therefore, the simulation practitioner is forced to assume a coefficient of skewness of zero and a coefficient of kurtosis of three for all components of the random vector. The multivariate model of the following section, on the other hand, allows the modeling of each component with a flexible distribution.

3.2.2 Multivariate Johnson Translation System

Johnson (1949b) represents the joint distribution of the $k-$dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)'$ via the use of the $k-$dimensional normalizing translation

$$\mathbf{Z} = \gamma + \delta\mathbf{f}[\lambda^{-1}(\mathbf{X}-\xi)] \sim N_k(\mathbf{0}, \Sigma_\mathbf{Z}),$$

where $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)'$ is the $k-$dimensional standard normal random vector with mean $\mathbf{0}$ and $(k \times k)$ covariance matrix $\Sigma_\mathbf{Z}$. The transformation function $\mathbf{f}(\cdot)$ is defined as $\mathbf{f}(y_1, y_2, \ldots, y_k) = (f_1(y_1), f_2(y_2), \ldots, f_k(y_k))'$, where $f(\cdot)$ is one of the transformations introduced in Section 3.1. Furthermore, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)'$ and $\xi = (\xi_1, \xi_2, \ldots, \xi_k)'$ are the $k-$dimensional vectors of shape and location parameters, and $\delta = \mathbf{diag}(\delta_1, \delta_2, \ldots, \delta_k)'$ and $\lambda = \mathbf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)'$ are the diagonal matrices.

Given the k components' Johnson parameters $\mathbf{f}$, γ, δ, λ, ξ and the covariance matrix $\Sigma_\mathbf{Z}$, a random vector with the $k-$dimensional distribution can be obtained by first generating independent standard normal random variables V_i, $i = 1, 2, \ldots, k$, and performing the Cholesky decomposition on $\Sigma_\mathbf{Z}$; i.e.,

$\Sigma_{\mathbf{Z}} = \mathbf{PP}'$. Then, we let $\mathbf{Z} \equiv \mathbf{PV}$ where $\mathbf{V} = (V_1, V_2, \ldots, V_k)$ and apply the inverse transformation $\mathbf{X} = \xi + \lambda \mathbf{f}^{-1}[\delta^{-1}(\mathbf{Z} - \gamma)]$ to obtain the random vector $\mathbf{X}$ with the desired characteristic. McDaniel et al. (1988) discuss an application of the multivariate Johnson translation system to welfare policy analysis.

Stainfield et al. (1996) propose an alternative method of sampling a Johnson random vector $\mathbf{X}$ with mean vector μ and covariance matrix Σ. First, we generate a random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)'$ with independent standardized Johnson components satisfying $\mathrm{E}(Y_i) = 0$, $\mathrm{Var}(Y_i) = 1$, and $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j = 1, 2, \ldots, k$. Then, we identify the lower triangular Cholesky decomposition $\mathbf{L}$ of the covariance matrix Σ (i.e., $\mathbf{L} = \Sigma^{1/2}$), set σ to **diag** [$\mathrm{Var}^{1/2}(X_1)$, $\mathrm{Var}^{1/2}(X_2)$, …, $\mathrm{Var}^{1/2}(X_k)$], and obtain $\mathbf{x} = \mu + \sigma \mathbf{LY}$. This sampling method has been shown to always match the first three moments of the components and the covariance matrix of $\mathbf{X}$; however, it may fail to match the fourth moments of the components. Stainfield et al. (2004) use this distribution for developing a simulation model of a remanufacturing facility.

3.3 Input Processes Changing Over Time

Typical examples of input processes that change over time are (nonstationary) arrival processes whose rates vary over time; e.g., the arrivals of customers at a fast-food restaurant lunchtime, arrivals of telephone calls to a call center, and seasonal product demands. The common approach in modeling a stationary arrival process is to use a Poisson arrival process with rate λ, where the times between arrivals are independent and exponentially distributed with mean $1/\lambda$. A nonstationary arrival process is a generalization of the stationary arrival process, where the arrival rate $\lambda(t)$ is generalized to be a function of time t. The classical model used to represent nonstationary arrival processes is the nonhomogeneous Poisson process (NHPP) $\{N(t) : t \geq 0\}$, which is an arrival-counting process with $N(t)$ denoting the number of arrivals at time interval (0,t], and $\lambda(t)$ corresponding to the instantaneous arrival rate at time t. The corresponding mean-value function is given by

$$\mu(t) = \mathrm{E}[N(t)] = \int_0^t \lambda(k)dk \quad \text{for all } t \geq 0.$$

Several procedures have been developed for modeling and estimating the mean-value function (rate function) of an NHPP; see Kuhl et al. (2008) and Kuhl et al. (2010) and the references therein. In this tutorial, we consider the case where $\lambda(t)$ exhibits trends and cyclic effects and present the following parametric rate function proposed by Kuhl et al. (1997a) for capturing trends and cyclic effects.

$$\lambda(t) = \exp\left\{ \sum_{i=0}^{m} \alpha_i t^i + \sum_{k=1}^{p} \gamma_k \sin(w_k t + \phi_k) \right\}$$

The first term of this representation is the polynomial component that represents the long-term evolutionary trend, while the second term is the trigonometric rate function that represents the periodic effects. The exponent itself ensures that the instantaneous arrival rate is always positive. The public domain softwares *mp3ml3* and *mp3sim* developed by Kuhl et al. (1997a) can be used, respectively, for estimating and simulating this NHPP.

Kuhl et al. (1997b) and Kuhl and Wilson (2001) consider the more general case where the trigonometric rate function may exhibit asymmetric behavior or the periodic effects may exhibit nontrigonometric behavior. They propose a nonparametric approach called multiresolution procedure, which has been implemented by Kuhl et al. (2006) on a Web-based software. This procedure has been shown to be faster than the parametric procedure of Kuhl et al. (1997a).

4 CONCLUSION

In this tutorial, we first present techniques for fitting a standard distribution to the available data and assessing the goodness of the fit. We also consider the case in which standard input models may fall short of representing the key characteristic of the data, and present flexible input models, multivariate input models, and input models developed for processes changing over time. When appropriate, we provide the software that can be used for solving these input modeling problems.

REFERENCES

Banks, J., J. S. Carson, B. L. Nelson, and D. Nicol. 2001. *Discrete-Event System Simulation.* 3rd Edition. Upper Saddle River, New Jersey: Prentice Hall.

Biller, B. and B. L. Nelson. 2002. Answers to the top ten input modeling questions. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yucesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 35-40. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Biller, B. and S. Ghosh. 2006. Multivariate input processes. In *Handbooks in Operations Research and Management Science: Simulation*, ed. B. L. Nelson and S. G. Henderson. Elsevier Science, Amsterdam.

Biller, B. and C. Gunes. 2010. Solving high-dimensional simulation input-modeling problems. Submitted to *Surveys in Operations Research and Management Science*.

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A Guide to Simulation.* 2nd Edition. Springer-Verlag, New York.

DeBrota, D. J., R. S. Dittus, S. D. Roberts, J. R. Wilson, J. J. Swain, and S. Venkatraman. 1989. Modeling input processes with Johnson distributions. In *Proceedings of the 1989 Winter Simulation Conference*, ed. E. A. MacNair, K. J. Musselman, and P. Heidelberger, 308 – 318. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

DeBrota D. J., R. S. Dittus, S. D. Roberts, and J. R. Wilson. 1989a. Visual interactive fitting of bounded Johnson distributions. *Simulation* 52:199-205.

Hill, I. D., R. Hill, and R. L. Holder. 1976. Fitting johnson curves by moments. *Applied Statistictics* 25:180-189.

Johnson, N. L. 1949a. Systems of frequency curves generated by methods of translation. *Biometrika* 36:149-176.

Johnson, N. L. 1949b. Bivariate distributions based on simple translation systems. *Biometrika* 36:297-304.

Johnson, M. E. 1987. *Multivariate Statistical Simulation*. Wiley, New York.

Johnson, R. A. and D. W. Wichern. 2001. *Applied Multivariate Statistical Analysis.* 6th Edition. Prentice Hall, New Jersey.

Kuhl, M. E., J. R. Wilson, and M. A. Johnson. 1997a. Estimating and simulating Poisson processes having trends or multiple periodicities. *IEE Transactions* 29:201-211.

Kuhl, M. E., H. Damerdji, and J. R. Wilson. 1997b. Estimating and simulating Poisson processes with trends or asymmetric cyclic effects. In *Proceedings of the 1997 Winter Simulation Conference*, ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 287 – 295. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Kuhl, M. E. and J. R. Wilson. 2001. Modeling and simulating Poisson processes having trends or nontrigonometric cyclic effects. *European Journal of Operational Research* 133:566-582.

Kuhl, M. E., S. G. Sumant, and J. R. Wilson. 2006. An automated multiresolution procedure for modeling complex arrival processes. *INFORMS Journal on Computing* 18:3-18.

Kuhl, M. E., S. C. Deo, and J. R. Wilson. 2008. Smooth flexible models of nonhomogeneous Poisson processes using one or more process realizations. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. Hill, L. Moench, and O. Rose, 353 – 361. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Kuhl, M. E., N. M. Steiger, E. K. Lada, M. A. Wagner, and J. R. Wilson. 2009. Introduction to modeling and generating probabilistic input processes for simulation. In *Proceedings of the 2009 Winter Simulation Conference*, ed. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 184-202. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Kuhl, M. E., J. S. Ivy, E. K. Lada, N. M. Steiger, M. A. Wagner, and J. R. Wilson. 2010. Multivariate input models for stochastic simulation. In preparation for submission to *Journal of Simulation*.

Law, A. M. 2007. *Simulation Modeling and Analysis.* 4th Edition. New York: McGraw-Hill.

Leemis, L. M. 2004. Building credible input models. In *Proceedings of the 2004 Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 29-40. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

McDaniel, R. R., R. S. Sullivan, and J. R. Wilson. 1988. A simulation model for welfare policy analysis. *Socio-Economic Planning Sciences* 22:157-165.

Nelson, B. L. and M. Yamnitsky. 1998. Input modeling tools for complex problems. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 105–112. Piscataway, New Jersey: Institute of Electronic and Electronics Engineers.

Pearson, K. 1895. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London* 91:343-358.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Fannery. 2007. *Numerical recipies in C: The art of scientific computing.* 3rd Edition. Cambridge University Press.

Ramberg, J. S. and B. W. Schmeiser. 1974. An approximate method for generating asymmetric random variables. *Communications of the Association for Computing Machinery* 17:78-82.

Schmeiser, B. and S. Deutsch. 1977. A versatile four-parameter family of probability distributions, suitable for simulation. *IIE Transactions* 9:176-182.

Slifker, J. F. and S. S. Shapiro. 1980. The Johnson System: Selection and parameter estimation. *Technometrics* 22:239-246.

Stanfield, P. M., J. R. Wilson, G. A. Mirka, N. F. Glasscock, J. P. Psihogios, and J. R. Davis. 1996. Multivariate input modeling with Johnson distributions. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 1457 – 1464. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Stanfield, P. M., J. R. Wilson, and R. E. King. 2004. Flexible modeling of correlated operation times with application in product-reuse facilities. *International Journal of Production Research* 42:2179-2196.

Swain, J. J., S. Venkatraman, and J. R. Wilson. 1988. Least-Squares estimation of distribution functions in Johnson's translation system. *Journal of Statistical Computation and Simulation* 29:271-297.

Tong, Y. L. 1990. *The Multivariate Normal Distribution*. New York: Springer-Verlag.

Vincent, S. 1998. Input data analysis. In the *Handbook of Simulation*, ed. J. Banks, 55 – 91. New York: John Wiley & Sons.

AUTHOR BIOGRAPHIES

BAHAR BILLER is an assistant professor of Operations Management and Manufacturing at Carnegie Mellon University. She received her PhD from Northwestern University. Her primary research interest lies in the area of computer simulation experiments for stochastic systems and more specifically, in the simulation methodology for dependent input processes with applications to financial markets and global supply chains.

CANAN GUNES is a PhD candidate in the Tepper School of Business at Carnegie Mellon University. Her research interests include the design of large-scale simulations with applications to inventory management, and the applications of Operations Research techniques (e.g., vehicle routing, inventory control) tailored to the problems of non-profit sectors.