Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

INTRODUCTION TO MODELING AND GENERATING PROBABILISTIC INPUT PROCESSES FOR SIMULATION

Michael E. Kuhl

Industrial & Systems Engineering Department
Rochester Institute of Technology
Rochester, NY 14623, U.S.A.

Julie S. Ivy

Edward P. Fitts Department of Industrial
and Systems Engineering
North Carolina State University
Raleigh, NC 27695, U.S.A.

Emily K. Lada

SAS Institute Inc.
100 SAS Campus Drive, R5413
Cary, NC 27513, U.S.A.

Natalie M. Steiger

Maine Business School
University of Maine
Orono, ME 04469, U.S.A.

Mary Ann Wagner

SAIC
7990 Science Applications Ct
Mail Stop CV-63
Vienna, VA 22182, U.S.A.

James R. Wilson

Edward P. Fitts Department of Industrial
and Systems Engineering
North Carolina State University
Raleigh, NC 27695, U.S.A.

ABSTRACT

Techniques are presented for modeling, fitting, and generating many of the univariate probabilistic input processes that drive discrete-event simulation experiments. Emphasis is given to the generalized beta distribution family, the Johnson translation system of distributions, and the Bézier distribution family because of the flexibility of these families to model a wide range of distributional shapes that arise in practical applications. Also discussed are nonparametric and semiparametric techniques for modeling and simulating time-dependent arrival streams using nonhomogeneous Poisson processes. Public-domain software implementations and current applications are presented for each input-modeling technique. The applications range from pharmaceutical manufacturing and medical decision analysis to smart-materials research and healthcare systems analysis. Many of the references include live hyperlinks providing online access to the referenced material.

1 INTRODUCTION

One of the main problems in the design and construction of stochastic simulation experiments is the selection of valid input models—i.e., probability distributions that accurately mimic the behavior of the random input processes driving the system under study. Often the following interrelated difficulties arise in attempts to use standard distribution families for simulation input modeling:

1. Standard distribution families cannot adequately represent the probabilistic behavior of many real-world input processes, especially in the tails of the underlying distribution.
2. The parameters of the selected distribution family are troublesome to estimate from either sample data or subjective information (expert opinion).
3. Fine-tuning or editing the shape of the fitted distribution is difficult because (i) there are a limited number of parameters available to control the shape of the fitted distribution, and (ii) there is no effective mechanism for directly manipulating the shape of the fitted distribution while simultaneously updating the corresponding parameter estimates.

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

In modeling a simulation input process, the practitioner must identify an appropriate distribution family and then estimate the corresponding distribution parameters; and the problems enumerated above can hinder the progress of both of these model-building activities.

The conventional approach to identification of a stochastic simulation input model encompasses several procedures for using sample data to accept or reject each of the distribution families in a list of well-known alternatives. These procedures include (i) informal graphical techniques based on probability plots, frequency distributions, or box-plots; and (ii) statistical goodness-of-fit tests such as the Kolmogorov-Smirnov, chi-squared, and Anderson-Darling tests. For a detailed discussion of these procedures, see Sections 6.3–6.6 of Law (2007). Unfortunately, none of these procedures is guaranteed to yield a definitive conclusion. For example, identification of an input distribution can be based on visual comparison of superimposed graphs of a histogram of the available data set and the fitted probability density function (p.d.f.) for each of several alternative distribution families. In this situation, however, the final conclusion depends largely on the number of class intervals (also called bins or cells) in the histogram as well as the class boundaries; and a different layout for the histogram could lead the user to identify a different distribution family. Similar anomalies can occur in the use of statistical goodness-of-fit tests. In small samples, these tests can have very low power to detect lack of fit between the empirical distribution and each alternative theoretical distribution, resulting in an inability to reject any of the alternative distributions. In large samples, moreover, practically insignificant discrepancies between the empirical and theoretical distributions often appear to be statistically significant, resulting in rejection of all the alternative distributions.

After somehow identifying an appropriate family of distributions to model an input process, the simulation user also faces problems in estimating the associated distribution parameters. The user often attempts to match the mean and standard deviation of the fitted distribution with the sample mean and standard deviation of a data set, but shape characteristics such as the sample skewness and kurtosis are less frequently considered when estimating the parameters of an input distribution. Some estimation methods, such as maximum likelihood and percentile matching, may simply fail to yield parameter estimates for some distribution families. Even if several distribution families are readily fitted to a set of sample data, the user generally lacks a definitive basis for selecting the appropriate "best-fitting" distribution.

The task of building a simulation input model is further complicated if sample data are not available. In this situation, identification of an appropriate distribution family is arbitrarily based on whatever information can be elicited from knowledgeable individuals (experts); and the corresponding distribution parameters are computed from subjective estimates of simple numerical characteristics of the underlying distribution such as the mode, selected percentiles, or low-order moments. In summary, simulation practitioners lack a clear-cut, definitive procedure for identifying and estimating high-fidelity stochastic input models (or even merely acceptable, "rough-cut" input models); consequently, simulation output analysis is often based on input processes of questionable validity.

In this article, techniques are presented for modeling and generating the probabilistic input processes that drive many simulation experiments, with the primary focus on methods designed to alleviate many of the difficulties encountered in using conventional approaches to simulation input modeling. Univariate input models are discussed in §2, with emphasis on the generalized beta distribution family, the Johnson translation system of distributions, and the Bézier distribution family. Some nonparametric and semiparametric techniques for modeling and simulating time-dependent arrival streams are discussed in §3. Finally conclusions and recommendations are presented in §4. The slides for the oral presentation of this article are available online via <www.ise.ncsu.edu/jwilson/files/wsc09atim.pdf>. Kuhl et al. (2009a, 2009b) provide a more detailed discussion of the topics covered in this article.

2 UNIVARIATE INPUT MODELS

2.1 Generalized Beta Distribution Family

Suppose X is a continuous random variable with lower limit a and upper limit b whose distribution is to be approximated and then randomly sampled in a simulation experiment. In such a situation, it is often possible to model the probabilistic behavior of X using a generalized beta distribution, whose p.d.f. has the form

$$f_X(x) = \frac{\Gamma(\alpha_1+\alpha_2)(x-a)^{\alpha_1-1}(b-x)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)(b-a)^{\alpha_1+\alpha_2-1}} \text{ for } a \le x \le b\,, \tag{1}$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,\mathrm{d}t$ (for $z > 0$) denotes the gamma function. For graphs illustrating the wide range of distributional shapes achievable with generalized beta distributions, see pp. 92–93 of Hahn and Shapiro (1967); pp. 291–292 of Law (2007); or the slides accompanying this article as mentioned at the end of §1.

If X has the p.d.f. (1), then the cumulative distribution function (c.d.f.) of X is given by $F_X(x) = \Pr\{X \le x\} = \int_{-\infty}^x f_X(w)\,\mathrm{d}w$ for all real x. Unfortunately $F_X(\cdot)$ has no convenient analytical expression; but the mean and variance of X are respectively given by

$$\mu_X = \mathrm{E}[X] = \frac{\alpha_1 b + \alpha_2 a}{\alpha_1 + \alpha_2} \quad \text{and} \quad \sigma_X^2 = \mathrm{E}\big[\big(X - \mu_X\big)^2\big] = \frac{(b-a)^2\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}\,. \tag{2}$$

Recall that for a continuous p.d.f. $f_X(\cdot)$, a mode m is a local maximum of that function; and if there is a unique global maximum for $f_X(\cdot)$, then the p.d.f. is said to be unimodal, and m is usually called the “most likely value” of the random variable X. If $\alpha_1, \alpha_2 \ge 1$ and either $\alpha_1 > 1$ or $\alpha_2 > 1$, then the beta p.d.f. (1) is unimodal; and the mode is given by

$$m = \frac{(\alpha_1 - 1)b + (\alpha_2 - 1)a}{\alpha_1 + \alpha_2 - 2} \qquad (\alpha_1, \alpha_2 \ge 1 \text{ and } \alpha_1\alpha_2 > 1)\,. \tag{3}$$

Equation (2) reveals that key distributional characteristics of the generalized beta distribution are simple functions of the parameters a, b, α_1, and α_2; and this facilitates input modeling—especially in pilot studies in which rapid model development is critical.

Fitting Beta Distributions to Data or Subjective Information. Given a random sample $\{X_i : i = 1, \ldots, n\}$ of size n from the distribution to be estimated, let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote the order statistics obtained by sorting the $\{X_i\}$ in ascending order so that $X_{(1)} = \min\big\{X_i : i = 1, \ldots, n\big\}$ and $X_{(n)} = \max\big\{X_i : i = 1, \ldots, n\big\}$. We can fit a generalized beta distribution to this data set using the following sample statistics:

$$\widehat{a} = 2X_{(1)} - X_{(2)}\,, \quad \widehat{b} = 2X_{(n)} - X_{(n-1)}\,, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i\,, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \bar{X}\big)^2\,. \tag{4}$$

In particular the method of moment matching involves (i) setting the right-hand sides of the two parts of (2) equal to the sample mean $\bar{X}$ and the sample variance S^2, respectively; and (ii) solving the resulting equations for the corresponding estimates $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ of the shape parameters. In terms of the auxiliary quantities $d_1 = \big(\bar{X} - \widehat{a}\big)\big/\big(\widehat{b} - \widehat{a}\big)$ and $d_2 = S\big/\big(\widehat{b} - \widehat{a}\big)$, the moment-matching estimates of $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are given by

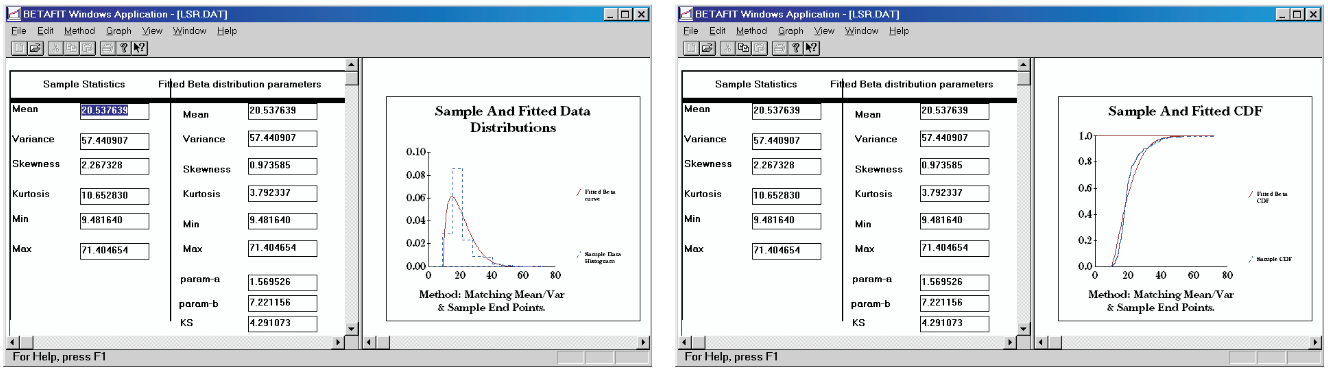
$$\widehat{\alpha}_1 = \frac{d_1^2(1-d_1)}{d_2^2} - d_1\,, \quad \widehat{\alpha}_2 = \frac{d_1(1-d_1)^2}{d_2^2} - (1-d_1)\,. \tag{5}$$

AbouRizk, Halpin, and Wilson (1994) discuss BetaFit, a Windows-based software package for fitting the generalized beta distribution to sample data by computing estimators $\widehat{a}$, $\widehat{b}$, $\widehat{\alpha}_1$, and $\widehat{\alpha}_2$ using the following estimation methods:

- moment matching with $\widehat{a} = X_{(1)}$ and $\widehat{b} = X_{(n)}$;
- feasibility-constrained moment matching, so that the feasibility conditions $\widehat{a} < X_{(1)}$ and $X_{(n)} < \widehat{b}$ are always satisfied;
- maximum likelihood (assuming a and b are known and thus are not estimated); and
- ordinary least squares (OLS) and diagonally weighted least squares (DWLS) estimation of the c.d.f.

Figure 1 demonstrates the application of BetaFit to a sample of 9,980 observations of end-to-end chain lengths (in angströms) of the ionic polymer Nafion based on the method of moment matching. Section 2.2 on the Johnson translation system of distributions provides further details on the origin of the Nafion data set and its relevance to the problem of predicting the stiffness properties of a certain class of smart materials. Like all the software packages mentioned in this article, BetaFit is in the public domain and is available on the Web via <www.ise.ncsu.edu/jwilson/page3>.

For rapid development of preliminary simulation models, practitioners often base an initial input model for the random variable X on subjective estimates $\widehat{a}$, $\widehat{m}$, and $\widehat{b}$ of the minimum, mode, and maximum, respectively, of the distribution of



Sample Statistics		Fitted Beta distribution parameters	
Mean	20.537639	Mean	20.537639
Variance	57.440907	Variance	57.440907
Skewness	2.267328	Skewness	0.973585
Kurtosis	10.652830	Kurtosis	3.792337
Min	9.481640	Min	9.481640
Max	71.404654	Max	71.404654
		param-a	1.569526
		param-b	7.221156
		KS	4.291073

Sample Statistics		Fitted Beta distribution parameters	
Mean	20.537639	Mean	20.537639
Variance	57.440907	Variance	57.440907
Skewness	2.267328	Skewness	0.973585
Kurtosis	10.652830	Kurtosis	3.792337
Min	9.481640	Min	9.481640
Max	71.404654	Max	71.404654
		param-a	1.569526
		param-b	7.221156
		KS	4.291073

Figure 1: Beta p.d.f. (left panel) and c.d.f. (right panel) fitted to 9,980 Nafion chain lengths using BetaFit

X. Although the triangular distribution is often used in such circumstances, it can yield excessively heavy tails—and hence grossly unrealistic simulation results—when the distance $\widehat{b}-\widehat{m}$ between the estimates of the upper limit and mode is much larger than the distance $\widehat{m}-\widehat{a}$ between the estimates of the mode and lower limit, or *vice versa*. The generalized beta distribution is usually a better choice in such situations, but there is some difficulty in selecting the shape parameters to yield the desired value $\widehat{m}$ for the mode.

In many project-management and quality-control applications, it is convenient to assume that the standard deviation of the random variable at hand is one-sixth of the corresponding range; and if the subjective estimates $\big(\widehat{b}-\widehat{a}\big)^2\big/36$ and $\widehat{m}$ of the variance and mode of X are respectively equated with the corresponding expressions for the variance and mode of a generalized beta distribution given in (2) and (3), then we must solve a cubic equation to obtain the corresponding shape parameters of the beta p.d.f. (1). In terms of the auxiliary quantity $q=\big(\widehat{m}-\widehat{a}\big)\big/\big(\widehat{b}-\widehat{a}\big)$, we see that in the special cases in which $q=0$ or $q=1$, the required shape parameters are exactly given by

$$\left.\begin{array}{l}\widehat{\alpha}_1 = 1 \;\text{ and }\; \widehat{\alpha}_2 = 3.87227 \;\text{ if } q = 0\,,\\ \widehat{\alpha}_1 = 3.87227 \;\text{ and }\; \widehat{\alpha}_2 = 1 \;\text{ if } q = 1\,.\end{array}\right\} \tag{6}$$

(For a detailed justification of (6), see the Appendix of Kuhl et al. (2009a), which contains exact computing formulas for the shape parameters of a beta distribution with user-specified values of the endpoints, mode, and variance.)

For the usual case in which $0<q<1$, remarkably accurate, simple approximations to the shape parameters of the beta distribution with minimum $\widehat{a}$, mode $\widehat{m}$, maximum $\widehat{b}$, and standard deviation $\big(\widehat{b}-\widehat{a}\big)\big/6$ can be conveniently calculated from the "asymmetry ratio" $r=\big(\widehat{b}-\widehat{m}\big)\big/\big(\widehat{m}-\widehat{a}\big)=(1-q)/q$ so that the required shape parameters are given by

$$\widehat{\alpha}_1 = \frac{r^2+3r+4}{r^2+1} \quad\text{and}\quad \widehat{\alpha}_2 = \frac{4r^2+3r+1}{r^2+1} \quad \text{if } \; 0<q<1\,; \tag{7}$$

see pp. 202–203 of Wilson et al. (1982). If $0.1 \le q \le 0.9$, then the error in the approximation (7) is less than 1.2%. To handle situations in which the estimated mode $\widehat{m}$ is very close to one of the estimated endpoints $\widehat{a}$ and $\widehat{b}$, see the Appendix of Kuhl et al. (2009a). In the application of beta distributions to a problem in medical decision making that is discussed briefly at the end of this section, the error in using the approximation (7) was essentially zero (that is, less than 10^{-8}) for each of 50 different beta distributions used in the associated simulation study; see also §2.4 of Kuhl et al. (2009a).

AbouRizk, Halpin, and Wilson (1991) discuss the Visual Interactive Beta Estimation System (VIBES), a Windows-based software package that enables graphically-oriented fitting of generalized beta distributions to subjective estimates of: (i) the endpoints a and b; and (ii) any of the following combinations of distributional characteristics—

- the mean μ_X and the variance σ_X^2,
- the mean μ_X and the mode m,
- the mode m and the variance σ_X^2,
- the mode m and an arbitrary quantile $x_p = F_X^{-1}(p)$ for $p \in (0,1)$, or
- two quantiles x_p and x_u for $p,u \in (0,1)$.

As a general-purpose tool for simulation input modeling, the generalized beta distribution family has the following advantages: (i) it is sufficiently flexible to represent with reasonable accuracy a wide diversity of distributional shapes; and (ii) its parameters are easily estimated from either sample data or subjective information. On the other hand, generating samples from the beta distribution is relatively slow; and in some applications, the time to generate beta random variables can be a substantial fraction of the overall simulation run time (Wilson et al. 1982).

Generating beta variates. Although most general-purpose simulation packages provide a generator of beta random variables, in our experience some care is required to verify the performance of a beta variate generator in cases where any shape parameter is less than one or is very large (say, greater than 30). Note that Equations (6)–(7) always yield $1 \le \alpha_1, \alpha_2 \le 4$ while equations (A1)–(A5) in the Appendix of Kuhl et al. (2009a) always yield $\alpha_1, \alpha_2 \ge 1$; and in these situations, we have obtained excellent results using two procedures available in Press *et al* (2007). To generate a generalized beta random variable X with minimum a, maximum b, and shape parameters α_1 and α_2, the first method uses `Gammadev` of Press *et al* (2007) to generate $Y(\alpha_1, \alpha_2)$, a standard beta random variable on the unit interval $[0, 1]$ with shape parameters α_1 and α_2; and then the desired random sample is given by

$$X = a + (b - a)Y(\alpha_1, \alpha_2)\,.$$

In terms of the incomplete beta function $I_x(\alpha_1, \alpha_2) = \left\{\Gamma(\alpha_1 + \alpha_2)\big/[\Gamma(\alpha_1)\Gamma(\alpha_2)]\right\} \int_0^x t^{\alpha_1 - 1}(1 - t)^{\alpha_2 - 1}\,\mathrm{d}t$ for $0 \le x \le 1$ (which coincides with the c.d.f. $F_{Y(\alpha_1,\alpha_2)}(x) = \Pr\{Y(\alpha_1, \alpha_2) \le x\}$ of a standard beta random variable $Y(\alpha_1, \alpha_2)$ for $0 \le x \le 1$), the second method for generating X is based on inversion of the c.d.f. of X,

$$X = F_X^{-1}(U) = a + (b - a)F_{Y(\alpha_1,\alpha_2)}^{-1}(U) = a + (b - a)I_U^{-1}(\alpha_1, \alpha_2)\,,$$

where $U \sim$ Uniform$[0, 1]$ is a random number and we use the procedure `invbetai` of Press *et al* (2007) to obtain a highly accurate approximation to $I_x^{-1}(\alpha_1, \alpha_2)$ for all x in $[0, 1]$.

Application of Beta Distributions to Pharmaceutical Manufacturing. Pearlswig (1995) provides a good example of a pharmaceutical manufacturing simulation whose credibility depended critically on the use of appropriate input models. In this study of the estimated production capacity of a plant that had been designed but not yet built, the usual three time estimates $\left(\widehat{a}, \widehat{m}, \text{and } \widehat{b}\ \right)$ were obtained from the process engineer for each of the operations in manufacturing a certain type of effervescent tablet. Unfortunately extremely conservative (i.e., large) estimates were provided for the upper limit $\widehat{b}$ of each operation time; and when triangular distributions were used to represent batch-to-batch variation in actual processing times for each operation within each step of production, the resulting bottlenecks resulted in very low estimates of the probability of reaching a prespecified annual production level.

As in many simulation applications in which subjective estimates $\widehat{a}$, $\widehat{m}$, and $\widehat{b}$ are elicited from experts, the estimate $\widehat{m}$ of the modal (most likely) time to perform a given operation was substantially more reliable than the estimates $\widehat{a}$ and $\widehat{b}$ of the lower and upper limits on the same operation time. When all the triangular distributions in the simulation were replaced by generalized beta distributions using (7) to ensure conformance to the engineer's estimate of the most likely processing time for each operation within each step, the resulting annual tablet production was in excellent agreement with the production of similar plants already in existence. This simple remedy restored the faith of management in the validity of the overall simulation model, which was subsequently used to finalize certain aspects of the design and operation of the new plant.

Application of Beta Distributions to Medical Decision Making. Cost-effectiveness analyses are frequently used in medical decision making for comparing various treatment or intervention alternatives. These studies involve uncertainty and random variability with respect to utility (i.e., effectiveness), probability, and cost estimates for disease states and interventions. There is variability between patients and parameter uncertainty, each reflected in the standard errors associated with simulation-based estimates of mean performance—for example, the expected values of the costs, quality-adjusted life years, and utilities resulting from alternative treatments. Therefore an accurate assessment of cost effectiveness must involve sensitivity analysis and must attempt to model the inherent variability and uncertainty in these parameter estimates. Probabilistic sensitivity analysis is one method for performing a multiway sensitivity analysis in which all parameters subject to uncertainty are varied simultaneously by Monte Carlo sampling from the distributions postulated for those parameters.

Xu et al. (2009) develop a decision-tree model for determining the cost effectiveness of cesarean delivery upon maternal request (CDMR) for women having a single childbirth without indications. Their model compares CDMR with trial of labor (TOL) considering all possible short- and long-term outcomes and the resulting consequences for the mother and neonate. This results in a decision tree containing over 100 chance events. For each parameter in their decision model, Xu et al. use either literature-based or expert opinion–based estimates for the mode, minimum, and maximum values. Typically there is limited information available for parameter distribution estimation; moreover, there is significant variability in the parameter values because of substantial uncertainty regarding mode of delivery with respect to utility measures, the probabilities of outcomes, and outcome costs. Extending the analysis of Xu et al., Kuhl et al. (2009a) fit beta distributions for all the utility and probability parameter estimates in the decision tree by two different approaches:

- Using the approximation based on Equations (6) and (7); and
- Using the version of the so-called "Beta PERT" distribution that is implemented in the `@RISK` software (Palisade Corporation 2009) and that is usually termed the RiskPert distribution.

Kuhl et al. (2009a) discuss the similarities and differences in the results based on both approaches to the use of beta distributions in this application. For some scenarios postulated by Kuhl et al., there was a significant difference in the effectiveness of CDMR and TOL (i.e., the 95% confidence interval for the mean difference in the utility between CDMR and TOL did not include zero) when using beta distributions fitted by each method. In other scenarios, there was a significant difference in the effectiveness of CDMR and TOL only when using beta distributions fitted via Equations (6) and (7). Finally in all other scenarios considered by Kuhl et al., the difference in effectiveness of CDMR and TOL was not significant for either method of fitting beta distributions.

The primary disadvantage of using the RiskPert distribution (and other versions of the "Beta PERT" distribution) is that its variance is a function of its mean and mode; and thus the user cannot specify the variance of a fitted beta distribution independently of its mean or mode. In some applications, it may be necessary to study systematically the sensitivity of the simulation-generated results to changes in the assumed values of the mode and variance of each input random variable; and in this case the development given in the Appendix of Kuhl et al. (2009a) can be used to investigate the impact of independently varying the postulated values of the mode and variance of the fitted beta distribution.

2.2 Johnson Translation System of Distributions

Starting from a continuous random variable X whose distribution is unknown and is to be approximated and subsequently sampled, Johnson (1949) proposes the idea of inferring an appropriate distribution by identifying a suitable "translation" (or transformation) of X to a standard normal random variable Z with mean 0 and variance 1 so that $Z \sim N(0,1)$. The translations have the form

$$Z = \gamma + \delta \cdot g\left(\frac{X-\xi}{\lambda}\right), \tag{8}$$

where γ and δ are shape parameters, λ is a scale parameter, ξ is a location parameter, and $g(\cdot)$ is a function whose form defines the four distribution families in the Johnson translation system,

$$g(y) = \begin{cases} \ln(y), & \text{for } S_L \text{ (lognormal) family,} \\ \ln\Big(y + \sqrt{y^2+1}\Big), & \text{for } S_U \text{ (unbounded) family,} \\ \ln[y/(1-y)], & \text{for } S_B \text{ (bounded) family,} \\ y, & \text{for } S_N \text{ normal family.} \end{cases}$$

DeBrota et al. (1989a) detail the advantages of the Johnson translation system of distributions for simulation input modeling, especially in comparison with the triangular, beta, and normal distribution families.

Johnson Distribution and Density Functions. If (8) is an exact normalizing translation of X to a standard normal random variable, then the c.d.f. of X is given by

$$F_X(x) = \Phi\Big[\gamma + \delta \cdot g\Big(\frac{x-\xi}{\lambda}\Big)\Big] \quad \text{for all } x \in \mathcal{H},$$

where: (i) $\Phi(z) = (2\pi)^{-1/2} \int_{-\infty}^{z} \exp\bigl(-\tfrac{1}{2}w^2\bigr)\,\mathrm{d}w$ denotes the c.d.f. of the $N(0,1)$ distribution; and (ii) the space $\mathcal{H}$ of X is

$$\mathcal{H} = \begin{cases} [\xi, +\infty), & \text{for } S_L \text{ (lognormal) family,} \\ (-\infty, +\infty), & \text{for } S_U \text{ (unbounded) family,} \\ [\xi, \xi + \lambda], & \text{for } S_B \text{ (bounded) family,} \\ (-\infty, +\infty), & \text{for } S_N \text{ normal family.} \end{cases}$$

The p.d.f. of X is given by

$$f_X(x) = \frac{\delta}{\lambda(2\pi)^{1/2}} g'\left(\frac{x-\xi}{\lambda}\right) \exp\left\{-\frac{1}{2}\left[\gamma + \delta \cdot g\left(\frac{x-\xi}{\lambda}\right)\right]^2\right\}$$

for all $x \in \mathcal{H}$, where

$$g'(y) = \begin{cases} 1/y, & \text{for } S_L \text{ (lognormal) family,} \\ 1\big/\sqrt{y^2+1}, & \text{for } S_U \text{ (unbounded) family,} \\ 1/[y(1-y)], & \text{for } S_B \text{ (bounded) family,} \\ 1, & \text{for } S_N \text{ normal family.} \end{cases}$$

For graphs illustrating the diversity of distributional shapes that can be achieved with the Johnson system of univariate distributions, see DeBrota (1989a) or the slides accompanying this article as mentioned at the end of §1.

Fitting Johnson Distributions to Sample Data. The process of fitting a Johnson distribution to sample data involves first selecting an estimation method and the desired translation function $g(\cdot)$ and then obtaining estimates of the four parameters γ, δ, λ, and ξ. The Johnson translation system of distributions has the flexibility to match (i) any feasible combination of values for the mean μ_X, variance σ_X^2, skewness $\mathrm{Sk}_X = \mathrm{E}\bigl[(X-\mu_X)^3/\sigma_X^3\bigr]$, and kurtosis $\mathrm{Ku}_X = \mathrm{E}\bigl[(X-\mu_X)^4/\sigma_X^4\bigr]$; or (ii) sample estimates of the moments μ_X, σ_X^2, Sk_X, and Ku_X. Moreover, in principle the skewness Sk_X and kurtosis Ku_X uniquely identify the appropriate translation function $g(\cdot)$. Although there are no closed-form expressions for the parameter estimates based on the method of moment matching, these quantities can be accurately approximated using the iterative procedure of Hill, Hill, and Holder (1976). Other estimation methods may also be used to fit Johnson distributions to sample data—for example, in the FITTR1 software package (Swain, Venkatraman, and Wilson 1988), the following methods are available:

- OLS and DWLS estimation of the c.d.f.;
- minimum L_1 and L_∞ norm estimation of the c.d.f.;
- moment matching; and
- percentile matching.

Fitting S_B Distributions to Subjective Information. DeBrota et al. (1989b) discuss VISIFIT, a public-domain software package for fitting Johnson S_B distributions to subjective information, possibly combined with sample data. The user must provide estimates of the endpoints a and b together with any two of the following characteristics:

- the mode m;
- the mean μ_X;
- the median $x_{0.5}$;
- arbitrary quantile(s) x_p or x_u for $p, u \in (0,1)$;
- the width of the central 95% of the distribution; or
- the standard deviation σ_X.

Generating Johnson Variates by Inversion. After a Johnson distribution has been fitted to a data set, generating samples from the fitted distribution is straightforward. First, a standard normal variate $Z \sim N(0,1)$ is generated. Then the corresponding realization of the Johnson random variable X is found by applying to Z the inverse translation

$$X = \xi + \lambda \cdot g^{-1}\left(\frac{Z-\gamma}{\delta}\right),$$

where for all real z we define the inverse translation function

$$g^{-1}(z) = \begin{cases} e^{z}, & \text{for } S_L \text{ (lognormal) family,} \\ \left(e^{z} - e^{-z}\right)/2, & \text{for } S_U \text{ (unbounded) family,} \\ 1/\left(1 + e^{-z}\right), & \text{for } S_B \text{ (bounded) family,} \\ z, & \text{for } S_N \text{ (normal) family.} \end{cases}$$

Although most popular general-purpose simulation packages provide an acceptable generator of standard normal random variables, we are particularly interested in generating Z by the method of inversion, $Z = \Phi^{-1}(U)$, where $U \sim \text{Uniform}[0,1]$ is a random number and we use the approximation to $\Phi^{-1}(\cdot)$ that is available via `Normaldist` of Press *et al* (2007). Also recommended is the approximation to $\Phi^{-1}(\cdot)$ given in Section 26.2.22 of Abramowitz and Stegun (1972).

Application of Johnson Distributions to Smart Materials Research. Matthews et al. (2006) and Weiland et al. (2005) present a multiscale modeling approach for the prediction of material stiffness of a certain class of smart materials called ionic polymers. The material stiffness depends on multiple parameters, including the effective length of the polymer chains composing the material. In a case study of Nafion, a specific type of ionic polymer, Matthews et al. (2006) develop a simulation model of the conformation of Nafion polymer chains on a nanoscopic level, from which a large number of end-to-end chain lengths are generated. The p.d.f. of end-to-end distances is then estimated and used as an input to a macroscopic-level mathematical model to quantify material stiffness.

Figure 2 shows the empirical distribution of 9,980 simulation-generated observations of end-to-end Nafion chain lengths (in angströms). Superimposed on the empirical distribution is the result of using the DWLS estimation method to fit an unbounded Johnson (S_U) distribution to the chain length data. Figure 2 reveals a remarkably accurate fit to the given data set. Furthermore, comparing the Johnson fit in Figure 2 with the beta fits for the same data set in Figure 1, we see that the Johnson distribution is able to capture certain key aspects of the Nafion data set that the beta distribution is unable to represent adequately.

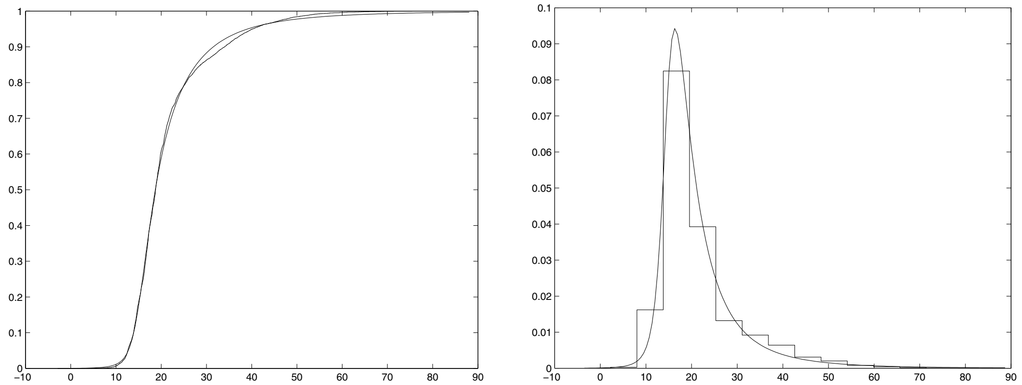


Figure 2: Johnson S_U c.d.f. (left panel) and p.d.f. (right panel) fitted to 9,980 Nafion chain lengths using FITTR1

Matthews et al. (2006) and Weiland et al. (2005) conclude that the estimates of the p.d.f. of chain lengths obtained by fitting an appropriate Johnson p.d.f. to the data are more intuitive than those using other density estimation techniques for the following reasons. First, it is possible to write down an explicit functional form for the Johnson p.d.f. $f_X(x)$ that is simple to differentiate. This is a crucial property since the second derivative $f_X''(x)$ of the p.d.f. will be used as an input to a mathematical model to estimate material stiffness. Second, there is a relatively simple relationship between the Johnson parameters and the material stiffness. Weiland et al. (2005) summarize the results of a sensitivity analysis for the Johnson parameters and the corresponding effect on material stiffness. In general, they find that increasing the location parameter ξ

leads to an increase in predicted stiffness. Similarly, increasing the shape parameter δ or decreasing the scale parameter λ both lead to marginally higher predicted material stiffness. Establishing a consistent relationship between these parameters and stiffness would extend the current theory to stiffness predictions, and it may ultimately be a step toward the custom design of materials with specific stiffness properties.

Application of Johnson Distributions to Healthcare. In a study of the arrival patterns of patients who have scheduled appointments at a community healthcare clinic, Alexopoulos et al. (2008) find that patient tardiness (i.e., the patient's deviation from the scheduled appointment time) is most accurately modeled using S_U distributions. Specifically they consider patient-tardiness data collected by the Partnership of Immunization Providers, a public-private collaboration involving the San Diego School of Medicine in the University of California together with community clinics and private-provider practices. Alexopoulos et al. (2008) perform an exhaustive analysis of 18 continuous distributions, and they conclude that the S_U distribution provides superior fits to the available data.

2.3 Bézier Distribution Family

Definition of Bézier Curves. In computer graphics, a Bézier curve is often used to approximate a smooth (continuously differentiable) function on a bounded interval by forcing the Bézier curve to pass in the vicinity of selected *control points* $\{\mathbf{p}_i \equiv (x_i, z_i)^{\mathrm{T}} : i = 0, 1, \ldots, n\}$ in two-dimensional Euclidean space. (In this article, all vectors are column vectors unless otherwise stated; and the roman superscript $^{\mathrm{T}}$ denotes the transpose of a vector.) Formally, a Bézier curve of degree n with control points $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n\}$ is given parametrically by

$$\mathbf{P}(t) = \sum_{i=0}^{n} B_{n,i}(t)\,\mathbf{p}_i \quad \text{for} \quad t \in [0,\ 1], \tag{9}$$

where $\mathbf{x} \equiv (x_0, x_1, \ldots, x_n)^{\mathrm{T}}$ and $\mathbf{z} \equiv (z_0, z_1, \ldots, z_n)^{\mathrm{T}}$, and where the *blending function* $B_{n,i}(t)$ (for all $t \in [0, 1]$) is the Bernstein polynomial

$$B_{n,i}(t) \equiv \frac{n!}{i!\,(n-i)!}\, t^i (1-t)^{n-i} \ \text{for}\ i = 0, 1, \ldots, n. \tag{10}$$

We let $P_x(t; n, \mathbf{x})$ and $P_z(t; n, \mathbf{z})$ respectively denote the abscissa and ordinate of $\mathbf{P}(t)$ for $t \in [0, 1]$.

Bézier Distribution and Density Functions. If X is a continuous random variable whose space is the bounded interval $[a,\ b]$ and if X has c.d.f. $F_X(\cdot)$, and p.d.f. $f_X(\cdot)$, then in principle we can approximate $F_X(\cdot)$ arbitrarily closely using a Bézier curve of the form (9) by taking a sufficient number $(n+1)$ of control points with appropriate values for the coordinates $(x_i, z_i)^{\mathrm{T}}$ of the ith control point $\mathbf{p}_i$ for $i = 0, \ldots, n$. If X is a Bézier random variable, then the c.d.f. of X is given parametrically by

$$\mathbf{P}(t) = \big\{x(t), F_X[x(t)]\big\}^{\mathrm{T}} \tag{11}$$

for $t \in [0,\ 1]$, where

$$x(t) = \sum_{i=0}^{n} B_{n,i}(t) x_i \quad \text{and} \quad F_X[x(t)] = \sum_{i=0}^{n} B_{n,i}(t) z_i\,. \tag{12}$$

Equation (12) reveals that the control points $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ constitute the parameters regulating all the properties of a Bézier distribution. Thus the control points must be arranged so as to ensure the basic requirements of a c.d.f.: (i) $F_X(x)$ is monotonically nondecreasing in the cutoff value x; (ii) $F_X(a) = 0$; and (iii) $F_X(b) = 1$. By utilizing the Bézier property that the curve described by (11)–(12) passes through the control points $\mathbf{p}_0$ and $\mathbf{p}_n$ exactly, we can ensure that $F_X(a) = 0$ if we take $\mathbf{p}_0 \equiv (a, 0)^{\mathrm{T}}$; and we can ensure that $F_X(b) = 1$ if we take $\mathbf{p}_n \equiv (b, 1)^{\mathrm{T}}$. See Wagner and Wilson (1996a) for a complete discussion of univariate Bézier distributions and their use in simulation input modeling.

If X is a Bézier random variable with c.d.f. $F_X(\cdot)$ given parametrically by (12), then it follows that the corresponding p.d.f. $f_X(x)$ for all real x is given parametrically by

$$\mathbf{P}^*(t) = \big\{x(t), f_X[x(t)]\big\}^{\mathrm{T}},$$

where $x(t)$ is given by (12) and

$$f_X[x(t)] = \frac{\sum_{i=0}^{n-1} B_{n-1,i}(t)\Delta z_i}{\sum_{i=0}^{n-1} B_{n-1,i}(t)\Delta x_i}$$

for $t \in [0,\ 1]$. In the last equation, we take $\mathbf{\Delta x} \equiv (\Delta x_0, \ldots, \Delta x_{n-1})^{\mathrm{T}}$ and $\mathbf{\Delta z} \equiv (\Delta z_0, \ldots, \Delta z_{n-1})^{\mathrm{T}}$, where $\Delta x_i \equiv x_{i+1} - x_i$ and $\Delta z_i \equiv z_{i+1} - z_i$ $(i = 0, 1, \ldots, n-1)$ represent the corresponding first differences of the x- and z-coordinates of the original control points $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n\}$ in the parametric representation (11) of the c.d.f.

Generating Bézier Variates by Inversion. The method of inversion can be used to generate a Bézier random variable whose c.d.f. has the parametric representation displayed in equations (11)–(12). Given a random number $U \sim$ Uniform[0, 1], we perform the following steps: (i) find $t_U \in [0,\ 1]$ such that

$$\sum_{i=0}^{n} B_{n,i}(t_U) z_i = U; \tag{13}$$

and (ii) deliver the variate

$$X = \sum_{i=0}^{n} B_{n,i}(t_U) x_i \,.$$

The solution to (13) can be computed by any root-finding algorithm such as Müller's method, Newton's method, or the bisection method. Codes to implement this approach to generating Bézier variates are available on Web site <www.ise.ncsu.edu/jwilson/page3>.

Using PRIME to Model Bézier Distributions. PRIME is a graphical, interactive software system that incorporates the methodology detailed in this section to help an analyst estimate the univariate input processes arising in simulation studies. PRIME is written entirely in the C programming language, and it has been developed to run under Microsoft Windows. A public-domain version of the software is available on the previously mentioned Web site. PRIME is designed to be easy and intuitive to use. The construction of a c.d.f. is performed through the actions of the mouse, and several options are conveniently available through menu selections. Control points are represented as small black squares, and each control point is given a unique label corresponding to its index i in equation (9). Figure 3 shows a typical session in PRIME, where the c.d.f. and p.d.f. windows are both displayed.

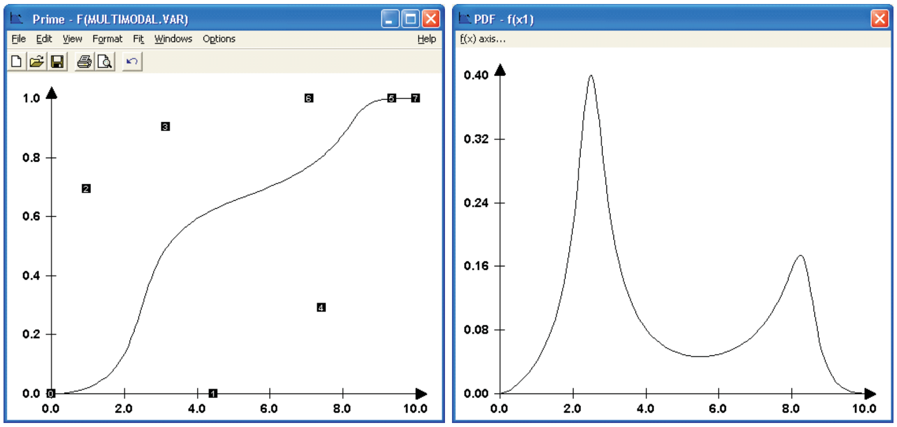


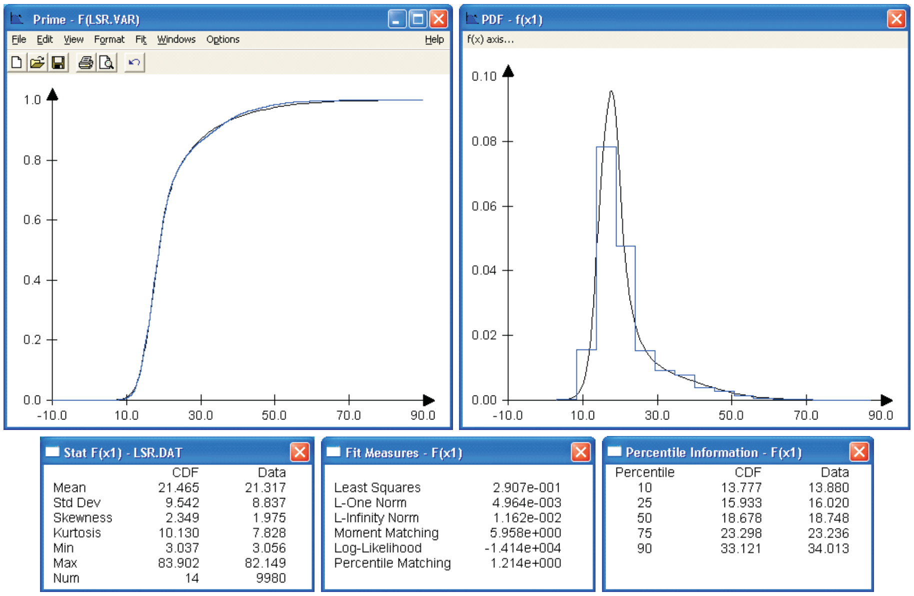
Figure 3: PRIME windows showing the Bézier c.d.f. (left panel) with its control points and the p.d.f. (right panel)

In the absence of data, PRIME can be used to model an input process conceptualized from subjective information or expertise. The representation of the conceptualized distribution is achieved by adding, deleting, and moving the control points via the mouse. Each control point acts like a "magnet" that pulls the curve in the direction of the control point, where the blending functions (i.e., the Bernstein polynomials defined by equation (10)) govern the strength of the "magnetic" attraction exerted on the curve by each control point. Selecting and dragging (i.e., moving) a control point causes the displayed c.d.f. to be updated (nearly) instantaneously. If they are displayed, the corresponding p.d.f., the first four moments (that is, the mean, variance, skewness, and kurtosis), and selected percentile values of the Bézier distribution are updated (nearly) simultaneously in adjacent windows so that the user gets immediate feedback on the effects of moving selected control points. Thus, the user has a variety of readily available indicators and measures, as well as visually appealing displays, to aid in the construction of the conceptualized distribution.

As detailed in Wagner and Wilson (1996a, 1996b), PRIME includes several standard estimation procedures for fitting distributions to sample data sets:

- OLS estimation of the c.d.f.;
- minimum L_1 and L_∞ norm estimation of the c.d.f.;
- maximum likelihood estimation (assuming a and b are known);
- moment matching; and
- percentile matching.

Figure 4 shows a Bézier distribution that was fitted to the same data set consisting of Nafion polymer chain lengths as shown in Figure 2. In this application of PRIME, we obtained the fitted Bézier distribution automatically, where: (i) the number of control points ($n+1=14$) was determined by the likelihood ratio test detailed in Wagner and Wilson (1996b); and (ii) the parameters $\mathbf{x}=(x_0,x_1,\ldots,x_n)^{\mathrm{T}}$ and $\mathbf{z}=(z_0,z_1,\ldots,z_n)^{\mathrm{T}}$ of the control points were estimated by the method of ordinary least squares. Figure 4 shows that a Bézier distribution yielded an excellent fit to the given data set.



	CDF	Data
Mean	21.465	21.317
Std Dev	9.542	8.837
Skewness	2.349	1.975
Kurtosis	10.130	7.828
Min	3.037	3.056
Max	83.902	82.149
Num	14	9980

Least Squares	2.907e-001
L-One Norm	4.964e-003
L-Infinity Norm	1.162e-002
Moment Matching	5.958e+000
Log-Likelihood	-1.414e+004
Percentile Matching	1.214e+000

Percentile	CDF	Data
10	13.777	13.880
25	15.933	16.020
50	18.678	18.748
75	23.298	23.236
90	33.121	34.013

Figure 4: Bézier distribution fitted to 9,980 Nafion chain lengths

The Bézier distribution family, which is entirely specified by its control points $\{\mathbf{p}_0,\mathbf{p}_1,\ldots,\mathbf{p}_n\}$, has the following advantages:

- It is extremely flexible and can represent a wide diversity of distributional shapes. For instance, Figure 3 depicts a multimodal distribution that is easily constructed using PRIME, yet impossible to achieve with other distribution families.
- If data are available, then the likelihood ratio test of Wagner and Wilson (1996a) can be used in conjunction with any of the estimation methods enumerated above to find automatically both the number and location of the control points.
- In the absence of data, PRIME can be used to determine the conceptualized distribution based on known quantitative or qualitative information that the user perceives to be pertinent.

- As the number $(n+1)$ of control points increases, so does the flexibility in fitting Bézier distributions. The interpretation and complexity of the control points, however, does not change with the number of control points.

Application of Bézier Distributions to Detection of Intrusions in Information Systems. In recent years, a large effort has been devoted to the development of procedures for rapidly and accurately detecting intrusions in information systems (Kim and Wilson 2006). Park (2005) derives event intensity (arrival-rate) data from log files generated by the Basic Security Module (BSM) of a Sun SPARC 10 workstation running the Solaris operating system and functioning as one of the components of the network simulated by the MIT Lincoln Laboratory; and he considers a Denial-of-Service (DoS) attack on the Sun workstation that leaves trails in the audit data. Park (2005) obtains event intensity data (that is, the number of events in successive one-second time intervals) derived from the BSM audit file for an observation period of 12,000 seconds on a specific day in the data sets from the MIT Lincoln Lab. Since the Sun system performs a specific routine for creating a log file every 60 seconds, this data set exhibits a repeated pattern every 60 seconds; see Figure 1 in Kim and Wilson (2006). After a careful analysis, Park separates this data set into the cyclic and noise parts as shown in Figures 2 and 3 of Kim and Wilson (2006).

For the detection of a DoS attack, the noise events must be monitored. Unfortunately the noise data are very sparse—in particular, only 60 of the 12,000 one-second time intervals contain noise events not related to the generation of a log file so that the estimated probability of occurrence of at least one noise event in a given one-second time interval is only 0.5%. No simple probabilistic models (in particular, the Poisson distribution) can provide an adequate fit to the observed noise data because of its high standard deviation. For the sample of 60 noise-event counts associated with one-second time intervals containing at least one noise event, the sample mean is 81 and the sample standard deviation is 154, which is almost twice as large as the mean. Park (2005) fits a Bézier distribution to the nonzero noise-event counts to drive a simulation-based performance evaluation of various intrusion-detection procedures. The fitted Bézier c.d.f. is displayed in Figure 5.

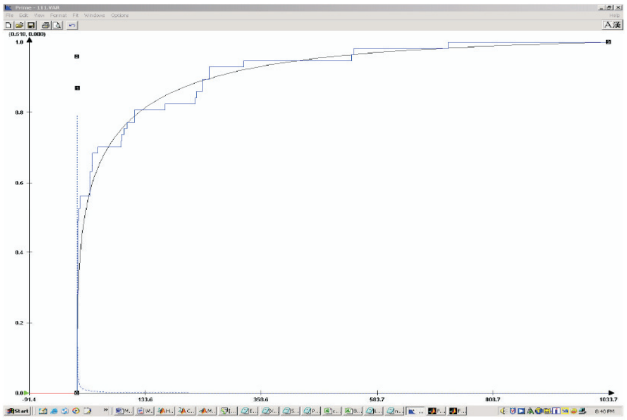


Figure 5: Bézier distribution fitted to 60 noise-event counts

3 TIME-DEPENDENT ARRIVAL PROCESSES

Time-varying arrival processes are routinely encountered in practical applications of industrial and systems engineering techniques. The following are typical situations in which the arrival rate of relevant entities depends strongly on time: demands for seasonal products such as lawn mowers; arrivals of patrons at an amusement park; arrivals of patients at an emergency room; and arrivals of telephone calls at a customer service center. To analyze or improve system operation in such situations, discrete-event stochastic simulation is often the technique of choice. Consequently, high-fidelity probabilistic input models are frequently needed to perform meaningful simulation experiments. In the past, nonhomogeneous Poisson processes (NHPPs) have been used successfully to model complex time-dependent arrival processes in a broad range of application domains (Lewis and Shedler 1976; Lee, Wilson, and Crawford 1991; and Pritsker et al. 1995).

An NHPP $\{N(t) : t \geq 0\}$ is a counting process such that $N(t)$ is the number of arrivals in the time interval $(0,t]$; and $\lambda(t)$, the instantaneous arrival rate at time t, is a nonnegative, integrable function satisfying the usual Poisson postulates so that the corresponding (cumulative) mean-value function is given by

$$\mu(t) \equiv \mathrm{E}[N(t)] = \int_0^t \lambda(z)\,\mathrm{d}z \quad \text{for all} \quad t \geq 0. \tag{14}$$

The rate or mean-value function of the NHPP $\{N(t) : t \geq 0\}$ completely characterizes the probabilistic behavior of the process.

Both parametric and nonparametric methods have been developed to estimate the rate or mean-value function of the process $\{N(t) : t \geq 0\}$ from observed arrival times. In this section we concentrate the discussion on a nonparametric approach of Leemis (1991, 2000, 2004); and we present the method in the context of a recent application to modeling and simulating unscheduled patient arrivals to a community healthcare clinic (Alexopoulos et al. 2008).

Suppose that we are given a time interval $(0, S]$ over which we observe several independent replications (realizations) of a stream of unscheduled patient arrivals, and that this stream of arrivals constitutes an NHPP with a time-dependent arrival rate $\lambda(t)$ for $t \in (0, S]$. For example, the observation interval might represent the time period within each weekday during which unscheduled patients may walk into a clinic—say, between 9:00 A.M. and 5:00 P.M. so that $S = 480$ min.

Suppose that k realizations of the arrival stream over this observation interval have been recorded so that we have n_i patient arrivals in the ith realization for $i = 1, 2, \ldots, k$; and thus we have a total of $n = \sum_{i=1}^{k} n_i$ patient arrivals accumulated over all realizations of the arrival stream. Moreover, let $\{t_{(i)} : i = 1, \ldots, n\}$ denote the overall set of arrival times for all unscheduled patients expressed as an offset from the beginning of the observation interval $(0, S]$ and then sorted in increasing order. Thus, for example, if we observed $n = 250$ patient arrivals over $k = 5$ days, each with an observation interval of length $S = 480$ min, then $t_{(1)} = 2.5$ min means that over all 5 days, the earliest patient arrival occurred 2.5 min after the clinic opened its doors to unscheduled arrivals on one of those days; and similarly, $t_{(2)} = 4.73$ min means that the second-earliest patient arrival occurred 4.73 min after the clinic opened its doors to unscheduled arrivals on one of those days.

Given that $\lambda(t)$ represents the rate of arrival of unscheduled patients for each time t in the observation interval $(0, S]$, we see that the mean-value function $\mu(t)$ representing the expected number of arrivals during the interval $(0, t]$ is given by (14). We take $t_{(0)} \equiv 0$ and $t_{(n+1)} \equiv S$ so that for $t_{(i)} < t \leq t_{(i+1)}$ and $i = 0, 1, \ldots, n$, a piecewise linear nonparametric estimator of $\mu(t)$ is

$$\widehat{\mu}(t) = \frac{in}{(n+1)k} + \left\{ \frac{n\big[t - t_{(i)}\big]}{(n+1)k\big[t_{(i+1)} - t_{(i)}\big]} \right\}; \tag{15}$$

see Leemis (1991). Figure 6 depicts the layout of $\widehat{\mu}(t)$.

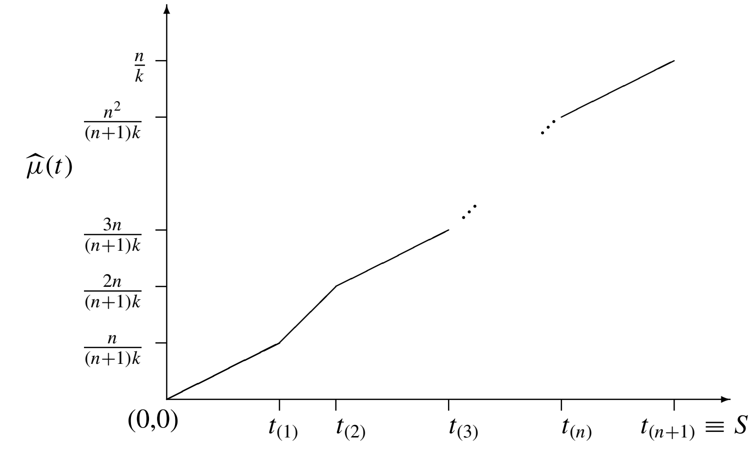


Figure 6: Nonparametric estimator of mean-value function

Equation (15) and Figure 6 provide a basis for modeling and simulating unscheduled patient-arrival streams when the arrival rate exhibits a strong dependence, for example, on the time of day.

To perform goodness-of-fit testing on the fitted mean-value function $\widehat{\mu}(t)$ for $t \in (0, S]$, we recommend the following cross-validation technique. Suppose that in addition to the realizations of the target arrival process that were used to compute the estimated mean-value function $\widehat{\mu}(t)$, we observe one additional realization $\left\{A_i' : i = 1, 2, \ldots, n'\right\}$ that is independent of the previously observed realizations, with the ith patient arriving at time A_i' for $i = 1, \ldots, n'$. If the target arrival stream is in fact an NHPP with true mean-value function $\mu(t)$ for $t \in (0, S]$, then the transformed arrival times $\left\{B_i' = \mu\left(A_i'\right) : i = 1, 2, \ldots, n'\right\}$ obtained by feeding each arrival time into the true mean-value function constitute a homogeneous Poisson process with an arrival rate of 1; and the corresponding transformed interarrival times $\left\{X_i' = B_i' - B_{i-1}' : i = 1, 2, \ldots, n'\right\}$ (with $B_0' \equiv 0$) constitute a random sample from an exponential distribution with a mean of 1.

It follows that an appropriate test for the adequacy of the fitted mean-value function $\widehat{\mu}(t)$ as an approximation to the true mean-value function $\mu(t)$ is to apply the Kolmogorov-Smirnov test to the data set $\left\{X_i'' = \widehat{\mu}\left(A_i'\right) - \widehat{\mu}\left(A_{i-1}'\right) : i = 1, 2, \ldots n'\right\}$ (with $A_0' \equiv 0$) consisting of estimates of the transformed interarrival times based on the estimated mean-value function, where the hypothesized c.d.f. in the goodness-of-fit test is $F_{X_i''}(x) = 1 - e^{-x}$ for all $x \geq 0$. For a comprehensive discussion of other techniques for assessing the goodness of fit of estimated arrival processes, see Lee, Wilson, and Crawford (1991); Kuhl, Wilson, and Johnson (1997); and Kuhl and Wilson (2000).

If the estimated mean-value function $\widehat{\mu}(t)$ passes the goodness-of-fit test outlined above, then we can use the simulation algorithm of Leemis (1991) as displayed in Figure 7 to generate a new stream of arrival times $\left\{A_i : i = 1, 2, \ldots\right\}$ over the time interval $(0, S]$ with approximately the same general pattern of dependence on time as in (15)—that is, with an arrival rate close to $\lambda(t)$ at each time t in the interval $(0, S]$.

```
[1] Set $i \leftarrow 1$ and $N \leftarrow 0$.
[2] Generate $U_i \sim \text{Uniform}(0, 1)$.
[3] Set $B_i \leftarrow -\ln(1 - U_i)$.
[4] While $B_i < n/k$ do
    Begin
        Set $m \leftarrow \left\lfloor \frac{(n+1)kB_i}{n} \right\rfloor$;
        Set $A_i \leftarrow t_{(m)} + \left\{t_{(m+1)} - t_{(m)}\right\}\left\{\frac{(n+1)kB_i}{n} - m\right\}$;
        Set $N \leftarrow N + 1$; Set $i \leftarrow i + 1$;
        Generate $U_i \sim \text{Uniform}(0, 1)$;
        Set $B_i \leftarrow B_{i-1} - \ln(1 - U_i)$.
    End
```

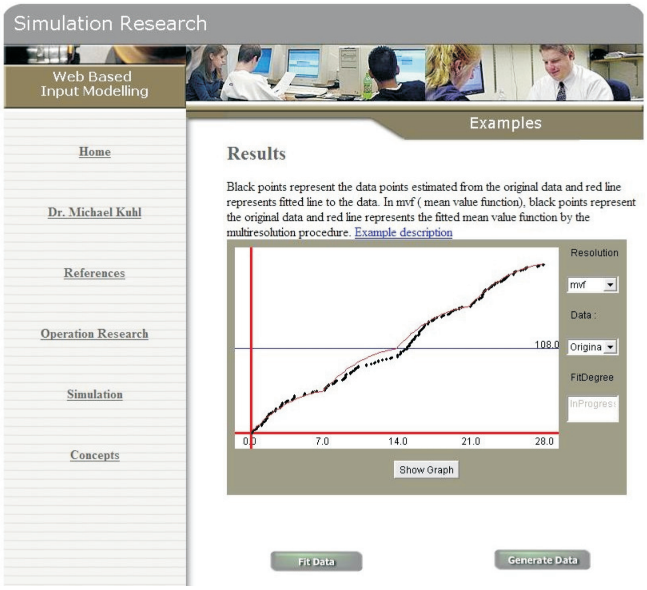
Figure 7: Algorithmic statement of the NHPP simulation procedure of Leemis (1991)

Note that in the simulation algorithm of Figure 7, $\lfloor z \rfloor$ denotes the greatest integer (or floor) function so that, for example, $\lfloor 3.7 \rfloor = 3$. Moreover, the total number of arrivals generated by this algorithm on one simulated realization of the arrival stream is given by the random variable N; and provided that $N > 0$, the ith patient will arrive at time A_i for $i = 1, \ldots, N$.

The main advantage of this approach to modeling and simulating time-dependent arrival processes is that it does not require the assumption of any particular functional form for the way in which the arrival rate $\lambda(t)$ depends on the time t since the beginning of the observation interval $(0, S]$. Moreover as $k \to \infty$, so that the number of realizations of the target arrival process becomes large, with probability 1 the estimated mean-value function $\widehat{\mu}(t)$ of equation (15) converges to the true mean-value function $\mu(t)$ for all $t \in (0, S]$. This means that the simulation algorithm given above (which is based on inversion of $\widehat{\mu}(t)$ so that $A_i = \widehat{\mu}^{-1}(B_i)$ for $i = 1, \ldots, N$) is also asymptotically valid as $k \to \infty$. For more information on this approach to modeling and simulation of time-dependent arrival processes, see Leemis (2004).

As an alternative to this nonparametric approach, Kuhl, Deo, and Wilson (2008) introduce a semiparametric method to model the mean-value function of the NHPP. The final estimate is obtained for the mean-value function of the following form: $\mu(t) = \mu(S)R(t)$ for $t \in [0, S]$, where $R(t)$ is a monotone increasing degree-r polynomial of the form

$$R(t) = \begin{cases} t/S, & \text{if } r = 1, \\ \displaystyle\sum_{k=1}^{r-1} \beta_k (t/S)^k + \left(1 - \sum_{k=1}^{r-1} \beta_k\right)(t/S)^r, & \text{if } r > 1. \end{cases}$$



Results

Black points represent the data points estimated from the original data and red line represents fitted line to the data. In mvf (mean value function), black points represent the original data and red line represents the fitted mean value function by the multiresolution procedure. Example description

Resolution: mvf

Data : Origina

FitDegree: InProgress

Figure 8: Web-based input-modeling software

This methodology can be used to fit an NHPP to one or more realizations of data from the process under study. The advantages that this method provides include a smooth function fit over the duration of the time interval without the need to specify the times of the arrival rate changes. For further details on the fitting procedures and data generation procedures, see Kuhl, Deo, and Wilson (2008).

Application to Organ Transplantation Policy Analysis. The United Network for Organ Sharing (UNOS) carried out a remarkable large-scale application of a simplified variant of this approach to modeling and simulating patient-arrival streams in the development and use of the UNOS Liver Allocation Model (ULAM) for analysis of the cadaveric liver-allocation system in the United States (see Harper et al. 2000). ULAM incorporated models of (a) the streams of liver-transplant patients arriving at 115 transplant centers, and (b) the streams of donated organs arriving at 61 organ procurement organizations in the United States—and virtually all these arrival streams exhibited strong dependencies on the time of day, the day of the week, and the season of the year as well as pronounced geographic effects.

Handling Arrival Processes Having Trends and Cyclic Effects. Kuhl and Wilson (2001) formulate a nonparametric method for modeling and simulating arrival processes that may exhibit a long-term trend or nested periodic phenomena (such as daily and weekly cycles), where the latter effects might not necessarily possess the symmetry of sinusoidal oscillations. Called a "multiresolution" procedure because of its ability to handle nested cyclic effects, this procedure has been implemented by Kuhl, Sumant, and Wilson (2006) in Web-based software, which is depicted in Figure 8 and is available online via <www.rit.edu/simulation>.

The procedure of Kuhl, Sumant, and Wilson (2006) involves the following steps at each resolution level corresponding to a basic cycle: (a) transforming the cumulative relative frequency of arrivals within the cycle (for example, the percentage of all arrivals as a function of the time of day within the daily cycle) to obtain a statistical model with approximately normal, constant-variance responses; (b) fitting a specially formulated polynomial to the transformed responses; (c) performing a likelihood ratio test to determine the degree of the fitted polynomial; and (d) fitting to the original (untransformed) responses a polynomial of the same form as in (b) with the degree determined in (c).

Kuhl, Sumant, and Wilson (2006) perform a comprehensive experimental performance evaluation to demonstrate the accuracy and flexibility of the automated multiresolution procedure. Figures 9 and 10 depict 90% tolerance bands for the underlying rate and mean-value functions, respectively, of an arrival process possessing one cyclic rate component and a long-term trend, where each tolerance band is based on applying the multiresolution procedure to 100 independent replications of the test process.

The inversion scheme of Kuhl and Wilson (2001) for simulating NHPPs fitted by the multiresolution estimation procedure is substantially faster than the corresponding inversion scheme of Kuhl, Wilson, and Johnson (1997) for simulating NHPPs

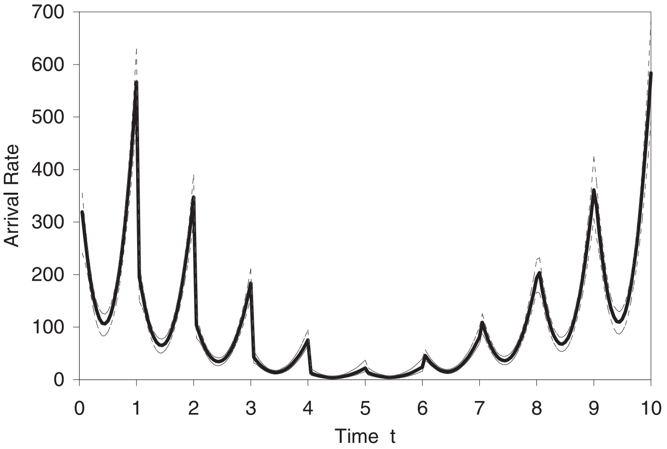


Figure 9: Fitted rate function over 100 replications of a test process with one cyclic rate component and long-term trend

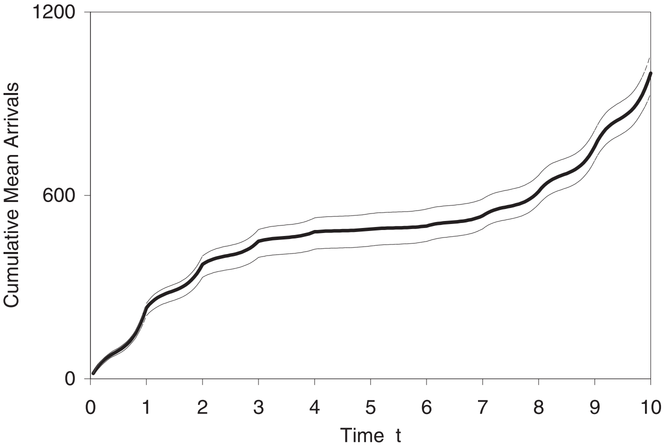


Figure 10: Fitted mean-value function over 100 replications of a test process with one cyclic rate component and long-term trend

that have a parametric rate function of the form

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

which is said to be of the type "Exponential-Polynomial-Trigonometric-with-Multiple-Periodicities" (EPTMP). Rate functions of type EPTMP were originally used in the UNOS Liver Allocation Model (Pritsker et al. 1995); and although the resulting fits were remarkably accurate, the times to generate realizations of the fitted NHPPs were too large in practice.

4 CONCLUSIONS AND RECOMMENDATIONS

The common thread running through this article is the focus on robust input models that are computationally tractable and sufficiently flexible to represent adequately many of the probabilistic phenomena that arise in many applications of discrete-event stochastic simulation. Specifically, input-modeling techniques are presented that deviate from conventional practice to produce realistic representations of the underlying input processes and hence yield greater fidelity in the resulting simulation output processes. In time, and with the growing understanding of the shortcomings of using classical distribution families in simulation experiments, we believe that these nonstandard techniques will in fact become the conventional procedures.

Notably missing from this article is a discussion of Bayesian and multivariate techniques for simulation input modeling, topics that we think will receive increasing attention from practitioners and researchers alike in the future. First we consider Bayesian techniques. In selecting the input models for a simulation, we must account for three main sources of uncertainty:

1. *Stochastic uncertainty* arises from dependence of the simulation output on the random numbers generated and used on each run—for example, the random number U used in generate a Bézier random variable X in (13), and the random numbers $\{U_i\}$ used to generate the arrival times $\{A_i\}$ in Figure 7.
2. *Model uncertainty* arises when the correct input model is unknown, and we must choose between alternative input models with different functional forms that adequately fit available sample data or subjective information—for example, the generalized beta, Johnson S_U, and Bézier distributions fitted to the Nafion data set as depicted in Figures 1, 2, and 4, respectively.
3. *Parameter uncertainty* arises when the parameters of the selected input model(s) are unknown and must be estimated from sample data or subjective information.

Although stochastic uncertainty is much more widely recognized by simulation practitioners than the other two types of uncertainty, it is not always a major source of variation in simulation output as demonstrated by Zouaoui and Wilson (2004) using an $M/G/1$ queueing system simulation in which stochastic uncertainty accounts for only 2% of the posterior variance of the average waiting time in the queue, while model uncertainty regarding the exact functional form of the service-time distribution accounts for 18% of the posterior variance—and thus 80% of the posterior variance is due to uncertainty regarding the exact numerical values of the arrival rate and the parameters of the service-time distribution. In such a situation, conventional approaches to input modeling have the potential to yield a grossly misleading picture of the inherent accuracy of simulation-generated system performance measures such as the average queue waiting time. For an introduction to Bayesian input modeling, see Chick (2001) and Zouaoui and Wilson (2003, 2004).

For some approaches to multivariate input modeling, see §3 of Kuhl et al. (2006) and §§2–4 of Kuhl et al. (2009b). Additional material on techniques for simulation input modeling will be posted to the Web site <www.ise.ncsu.edu/jwilson/more_info>.

REFERENCES

AbouRizk, S. M., D. W. Halpin, and J. R. Wilson. 1991. Visual interactive fitting of beta distributions. *Journal of Construction Engineering and Management* 117 (4): 589–605. Available online via <www.ise.ncsu.edu/jwilson/files/abourizk91jcem.pdf> [accessed May 29, 2009].

AbouRizk, S. M., D. W. Halpin, and J. R. Wilson. 1994. Fitting beta distributions based on sample data. *Journal of Construction Engineering and Management* 120 (2): 288–305. Available online via <www.ise.ncsu.edu/jwilson/files/abourizk94jcem.pdf> [accessed May 29, 2009].

Alexopoulos, C., D. Goldsman, J. Fontanesi, D. Kopald, and J. R. Wilson. 2008. Modeling patient arrival times in community clinics. *Omega* 36:33–43. Available online via <www.ise.ncsu.edu/jwilson/files/alex08omega.pdf> [accessed May 29, 2009].

Chick, S. E. 2001. Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research* 49 (5): 744–758.

DeBrota, D. J., R. S. Dittus, S. D. Roberts, J. R. Wilson, J. J. Swain, and S. Venkatraman. 1989a. 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. Available online via <www.ise.ncsu.edu/jwilson/files/wsc89jnsn.pdf> [accessed May 29, 2009].

DeBrota, D. J., R. S. Dittus, S. D. Roberts, and J. R. Wilson. 1989b. Visual interactive fitting of bounded Johnson distributions. *SIMULATION* 52 (5): 199–205. Available online via <www.ise.ncsu.edu/jwilson/files/debrota89sim .pdf> [accessed May 29, 2009].

Hahn, G. J., and S. S. Shapiro. 1967. *Statistical models in engineering*. New York: Wiley.

Harper, A. M., S. E. Taranto, E. B. Edwards, and O. P. Daily. 2000. An update on a successful simulation project: The UNOS liver allocation model. In *Proceedings of the 2000 Winter Simulation Conference*, ed. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 1955–1962. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online as <www.informs-sim.org/wsc00papers/267.pdf> [accessed May 29, 2009].

Hill, I. D, R. Hill, and R. L. Holder. 1976. Algorithm AS99: Fitting Johnson curves by moments. *Applied Statistics* 25 (2): 180–189.

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

Kim, S.-H., and J. R. Wilson. 2006. A discussion on 'Detection of intrusions in information systems by sequential change-point methods' by Tartakovsky, Rozovskii, Blažek, and Kim. *Statistical Methodology* 3:315–319. Available online via <www.ise.ncsu.edu/jwilson/files/kim06statmeth.pdf> [accessed May 29, 2009].

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, to appear. 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. 2009a. Univariate input models for stochastic simulation. *Journal of Simulation* in review. Available online via <www.ise.ncsu.edu/jwilson/files/kuhl09ajos>.pdf> [accessed July 14, 2009].

Kuhl, M. E., J. S. Ivy, E. K. Lada, N. M. Steiger, M. A. Wagner, and J. R. Wilson. 2009b. Multivariate input models for stochastic simulation. *Journal of Simulation* in review. Available online via <www.ise.ncsu.edu/jwilson/files/kuhl09bjos>.pdf> [accessed July 14, 2009].

Kuhl, M. E., E. K. Lada, N. M. Steiger, M. A. Wagner, and J. R. Wilson. 2006. Introduction to modeling and generating probabilistic input processes for simulation. In *Proceedings of the 2006 Winter Simulation Conference*, ed. L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 19–35. Available online via <www.informs-sim.org/wsc06papers/003.pdf> [accessed May 29, 2009].

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 (1): 3–18. Available online via <www.ise.ncsu.edu/jwilson/files/kuhl06joc.pdf> [accessed May 29, 2009].

Kuhl, M. E., and J. R. Wilson. 2000. Least squares estimation of nonhomogeneous Poisson processes. *Journal of Statistical Computation and Simulation* 67:75–108. Available online via <www.ise.ncsu.edu/jwilson/files/kuhl00jscs.pdf> [accessed May 29, 2009].

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 (3): 566–582. Available online via <www.ise.ncsu.edu/jwilson/files/kuhl01ejor.pdf> [accessed May 29, 2009].

Kuhl, M. E., J. R. Wilson, and M. A. Johnson. 1997. Estimating and simulating Poisson processes having trends or multiple periodicities. *IIE Transactions* 29 (3): 201–211. Available online via <www.ise.ncsu.edu/jwilson/files/kuhl97iie.pdf> [accessed May 29, 2009].

Law, A. M. 2007. *Simulation modeling and analysis*. 4th ed. New York: McGraw-Hill.

Lee, S., J. R. Wilson, M. M. Crawford. 1991. Modeling and simulation of a nonhomogeneous Poisson process having cyclic behavior. *Communications in Statistics—Simulation* 20:777–809. Available online via <www.ise.ncsu.edu/jwilson/files/lee91.pdf> [accessed May 29, 2009].

Leemis, L. M. 1991. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process. *Management Science* 37 (7): 886–900.

Leemis, L. M. 2000. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process from overlapping realizations. *Management Science* 46 (7): 989–998.

Leemis, L. M. 2004. Nonparametric estimation and variate generation for a nonhomogeneous Poisson process from event count data. *IIE Transactions* 36 (12): 1155–1160.

Lewis, P. A. W., and G. S. Shedler. 1976. Statistical analysis of non-stationary series of events in a data base system. *IBM Journal of Research and Development* 20:465–482.

Matthews, J. L., E. K. Lada, L. M. Weiland, R. C. Smith, and D. J. Leo. 2006. Monte Carlo simulation of a solvated ionic polymer with cluster morphology. *Smart Materials and Structures* 15 (1): 187–199. Available online via <www.ise.ncsu.edu/jwilson/files/matthews06sms.pdf> [accessed May 29, 2009].

Palisade Corp. 2009. Getting started in @RISK. Ithaca, New York: Palisade Corp. Available online via <www.palisade.com/risk/5/tips/EN/gs/> [accessed July 5, 2009].

Park, Y. 2005. *A statistical process control approach for network intrusion detection*. Ph.D. thesis, School of Industrial and Systems Engineering, Atlanta, Georgia. Available online via <hdl.handle.net/1853/6835> [accessed May 29, 2009].

Pearlswig, D. M. 1995. *Simulation modeling applied to the single pot processing of effervescent tablets.* Master's thesis, Integrated Manufacturing Systems Engineering Institute, North Carolina State University, Raleigh, North Carolina. Available online via <www.ise.ncsu.edu/jwilson/files/pearlswig95.pdf> [accessed May 29, 2009].

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 2007. *Numerical recipes: The art of scientific computing*. 3rd ed. Cambridge: Cambridge University Press.

Pritsker, A. A. B., D. L. Martin, J. S. Reust, M. A. Wagner, O. P. Daily, A. M. Harper, E. B. Edwards, L. E. Bennett, J. R. Wilson, M. E. Kuhl, J. P. Roberts, M. D. Allen, and J. F. Burdick. 1995. Organ transplantation policy evaluation. In *Proceedings of the 1995 Winter Simulation Conference*, ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, 1314–1323. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <www.ise.ncsu.edu/jwilson/files/pritsker95wsc.pdf> [accessed May 29, 2009].

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. Available online via <www.ise.ncsu.edu/jwilson/files/jnsn88jscs.pdf> [accessed May 29, 2009].

Wagner, M. A. F., and J. R. Wilson. 1996a. Using univariate Bézier distributions to model simulation input processes. *IIE Transactions* 28 (9): 699–711. Available online as <www.ise.ncsu.edu/jwilson/files/wagner96iie.pdf> [accessed May 31, 2008].

Wagner, M. A. F., and J. R. Wilson. 1996b. Recent developments in input modeling with Bézier distributions. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 1448–1456. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online as <www.ise.ncsu.edu/jwilson/files/wagner96wsc.pdf> [accessed May 29, 2009].

Weiland, L. M., E. K. Lada, R. C. Smith, and D. J. Leo. 2005. Application of rotational isomeric state theory to ionic polymer stiffness predictions. *Journal of Materials Research* 20 (9): 2443–2455. Available online via <www.ise.ncsu.edu/jwilson/files/weiland05jmr.pdf> [accessed May 29, 2009].

Wilson, J. R., D. K. Vaughan, E. Naylor, and R. G. Voss. 1982. Analysis of Space Shuttle ground operations. *Simulation* 38 (6): 187–203.

Xu, X., J. S. Ivy, D. A. Patel, S. N. Patel, D. G. Smith, S. B. Ransom, D. Fenner, and J. O. L. DeLancey. 2009. Lifelong pelvic floor consequences of cesarean delivery on maternal request for primigravid women with a single birth: A cost effectiveness analysis. *Journal of Women's Health* forthcoming. Available online via <www.ise.ncsu.edu/jwilson/files/xu09jwh.pdf> [accessed June 6, 2009].

Zouaoui, F., and J. R. Wilson. 2003. Accounting for parameter uncertainty in simulation input modeling. *IIE Transactions* 35 (3): 781–792. Available online via <www.ise.ncsu.edu/jwilson/files/zouaoui03iie.pdf> [accessed May 29, 2009].

Zouaoui, F., and J. R. Wilson. 2004. Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions* 36 (11): 1135–1151. Available online via <www.ise.ncsu.edu/jwilson/files/zouaoui04iie.pdf> [accessed May 29, 2009].

AUTHOR BIOGRAPHIES

MICHAEL E. KUHL is an associate professor in the Industrial and Systems Engineering Department at Rochester Institute of Technology. His e-mail address is <mekeie@rit.edu>.

JULIE S. IVY is an assistant professor in the Fitts Department of Industrial and Systems Engineering at NC State University. Her e-mail address is <jsivy@unity.ncsu.edu>.

EMILY K. LADA is an operations research development tester at the SAS Institute. Her e-mail address is <Emily.Lada@sas.com>.

NATALIE M. STEIGER is an associate professor of production and operations management in the University of Maine Business School. Her e-mail address is <nsteiger@maine.edu>.

MARY ANN WAGNER is a systems engineer at SAIC. Her e-mail address is <wagnermar@saic.com>.

JAMES R. WILSON is a professor in the Fitts Department of Industrial and Systems Engineering at NC State University. His e-mail address is <jwilson@ncsu.edu>.