Fitting Beta Distributions Based on Sample Data

By Simaan M. AbouRizk,[1] Associate Member, ASCE, Daniel W. Halpin,[2] Member, ASCE, and James R. Wilson[3]

Abstract: Construction operations are subject to a wide variety of fluctuations and interruptions. Varying weather conditions, learning development on repetitive operations, equipment breakdowns, management interference, and other external factors may impact the production process in construction. As a result of such interferences, the behavior of construction processes becomes subject to random variations. This necessitates modeling construction operations as random processes during simulation. Random processes in simulation include activity and processing times, arrival processes (e.g. weather patterns) and disruptions. In the context of construction simulation studies, modeling a random input process is usually performed by selecting and fitting a sufficiently flexible probability distribution to that process based on sample data. To fit a generalized beta distribution in this context, a computer program founded upon several fast, robust numerical procedures based on a number of statistical-estimation methods is presented. In particular, the following methods were derived and implemented: moment matching, maximum likelihood, and least-square minimization. It was found that the least-square minimization method provided better quality fits in general, compared to the other two approaches. The adopted fitting procedures have been implemented in BetaFit, an interactive, microcomputer-based software package, which is in the public domain. The operation of BetaFit is discussed, and some applications of this package to the simulation of construction projects are presented.

INTRODUCTION

In a wide diversity of construction simulation studies, it is often necessary to represent a particular sequence of simulation inputs as independent random variables taken from a common underlying probability distribution; and then the main objective of simulation input modeling is to approximate this distribution accurately and with a minimum of computational effort. A key element in making computer simulation accessible to construction practitioners is the automation of the complex statistical and numerical techniques required to achieve the desired accuracy within a simulation experiment.

Fitting statistical distribution to sample data can be found in many construction applications including risk analysis, quality control, and mostly, costing and scheduling. Al-Masri (1985) fitted lognormal and other distributions in studying earthmoving operations. Touran and Wiser (1992) discuss a Monte Carlo technique used for "range estimating" requiring the cost engineer to model the underlying distributions of unit cost data, for example. In their application, correlated data was to be modeled and their choice of distribution was lognormal. Additional examples include a recent

[1] Assoc. Prof., Civ. Engrg. Dept., Univ. of Alberta, Edmonton, Alberta, Canada T6G-2G7.

[2] Prof. and Head, Div. of Constr. Engrg. and Mgmt., Purdue Univ., West Lafayette, IN 47907.

[3] Prof., Dept. of Industrial Engrg., North Carolina State Univ., Box 7906, Raleigh, NC 27695.

Note. Discussion open until November 1, 1994. To extend the closing date one month, a written request must be filed with the ASCE Manager of Journals. The manuscript for this paper was submitted for review and possible publication on November 23, 1993. This paper is part of the ***Journal of Construction Engineering and Management***, Vol. 120, No. 2, June, 1994. ©ASCE, ISSN 0733-9364/94/0002-0288/$2.00 + $.25 per page. Paper No. 5156.

work by Farid and Aziz (1993), which involved fitting beta distributions to sample earthmoving data in the process of validating queuing models and simulating nonstationary travel times through the use of CYCLONE.

The main emphasis of the present application is input modeling for construction simulation. This includes modeling the variability of activity times for probabilistic project scheduling and work package cost variability in the context of range estimating. Construction simulation modeling (e.g. CYCLONE applications) also requires accurate statistical input modeling.

AbouRizk and Halpin (1992a) presented an empirical study showing that a flexible distribution is often required and is recommended to model activity times. AbouRizk et al. (1992) presented a method for modeling activity times in the absence of data for construction simulation. When a sample of data can be collected (e.g. from an on-going earthmoving operation), a proper distribution can be used to provide accurate representation of the underlying random process of the activity time or cost. To make this available for construction practitioners and researchers an automated technique that will read a collected sample of data and perform many numerical approximations to arrive at an appropriate beta distribution was developed.

DISTRIBUTION FITTING IN SIMULATION STUDIES

The conventional approach to simulation input modeling is to fit a probability distribution from a standard family of continuous distributions based on sample data. Compared to the discontinuities and irregularities of the cumulative density function (CDF) of the sample data set, this approach yields a smooth, regular approximation to the unknown CDF from which the sample was taken. This approach also eliminates the need for storing and manipulating large amounts of sample data when generating random samples from a particular input process.

The main difficulty in simulation input modeling is the broad range of distributional shapes that must be accommodated in practice. This motivates the use of a flexible family of distributions which is capable of attaining a wide variety of shapes. Among such families are the Pearson system (Johnson and Kotz 1970), the Johnson translation system (Johnson 1948), the Lambda distribution family and its modifications (Tukey 1960; Schmeiser and Deutch 1976; Ramberg et al. 1979), and the generalized beta family of distributions (Hahn and Shapiro 1967). In the present paper, the focus is on the generalized beta family of distributions, which has been widely used in a variety of construction engineering and management applications, is well known, and is available in virtually all simulation software packages.

The accuracy of distribution fitting in construction simulation greatly depends on the application involved and the statistical measure of performance required. A study by AbouRizk (1990) showed that if the parameter sought from the simulation is a mean measure of performance (e.g. mean project completion time or cost) the use of a triangular, lognormal, or beta distribution as input models in the simulation experiment could yield close values of the estimated mean as long as the means of the input models used are the same. As the order of the statistical estimate increases the simulation results begin reflecting more sensitivity to the distribution type used. For example, when the 90th percentile value of the waiting time in a truck queue is required, the use of a lognormal, triangular, or beta distributions within the model may yield considerably different results. Therefore, it is essential in such applications that the selected distribution to model the collected data truly reflects the properties of the data. This may be validated through common goodness of fit tests or by visual assessment of the fitted versus collected sample CDFs. A summary of the effect of input modeling on construction simulation can be found in AbouRizk and Halpin (1992b).

GENERALIZED BETA FAMILY OF DISTRIBUTIONS

The probability density function (PDF) for the generalized beta distribution defined on the interval $[L, U]$ with shape parameters a and b is given by

$$\begin{aligned} f(x) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{(x-L)^{a-1}(U-x)^{b-1}}{(U-L)^{a+b-1}} && \text{if } L \leq X \leq U \\ f(x) &= 0 && \text{otherwise} \end{aligned} \tag{1}$$

where the lower limit L and the upper limit U satisfy $L < U < \infty$; shape parameters a and b are positive; and gamma function $\Gamma(\cdot)$ is defined by

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt \qquad \text{for all } z > 0 \tag{2}$$

The corresponding CDF over the range $[L, U]$ with the shape parameters a and b is given by

$$\begin{aligned} F(x) &= 0 && \text{if } x < L \\ F(x) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_L^x \frac{(t-L)^{a-1}(U-t)^{b-1}}{(U-L)^{a+b-1}}\,dt && \text{if } L \leq x \leq U \\ F(x) &= 1 && \text{if } x > U \end{aligned} \tag{3}$$

If X is a random variable having the generalized beta distribution (3), then the mean, variance, skewness, and kurtosis of the X are respectively given by the following expressions:

$$\mu = E(X) = L + (U - L)\frac{a}{a+b} \tag{4}$$

$$\sigma^2 = E[(X-\mu)^2] = (U-L)^2 \frac{ab}{(a+b)^2(a+b+1)} \tag{5}$$

$$\alpha_3 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}} \tag{6}$$

and

$$\alpha_4 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{3(a+b+1)[2(a+b)^2 + ab(a+b-6)]}{ab(a+b+2)(a+b+3)} \tag{7}$$

see Hahn and Shapiro (pages 91–98, 126–128, 1967). Eqs. (6) and (7) show how the shape of the generalized beta density $f(x)$ is determined by the values of the two shape parameters a and b. Similarly, (4) and (5) show how the mean and variance of the generalized beta distribution depend on the location parameter L and the scale parameter $(U - L)$. Let $\Theta = (L, U, a, b)$ denote the vector of parameters defining the generalized beta density (1). To emphasize the dependence of the PDF, (1), and the CDF, (3), on this parameter vector, $f(x; \boldsymbol{\Theta})$ and $F(x; \boldsymbol{\Theta})$ will be subsequently used to represent these functions.

STATISTICAL METHODS FOR FITTING BETA DISTRIBUTIONS

As mentioned previously, the generalized beta probability distribution is frequently used in simulation studies to model the behavior of an input quantity that is subject to random variation or that is simply not known with certainty. Suppose that $\{X_i: 1 \leq i \leq n\}$ is a random sample of size n from the underlying probability distribution of interest. It is postulated that this unknown CDF has the form $F(x; \boldsymbol{\Theta})$ given by (3), and an accurate and computationally efficient statistical method to estimate $\boldsymbol{\Theta}$ to completely specify the required input model is being sought.

Moment Matching

To fit a postulated distribution with k unknown parameters based on sample data, the method of moment matching requires the following steps: (*a*) Compute the first k sample moments; (*b*) equate these statistics to the corresponding theoretical moments to obtain a system of k equations in the k unknown parameters; and (*c*) solve this equation system for the corresponding parameter estimates. Several variants of this basic idea have been specialized to the case of fitting beta distributions.

To fit a beta distribution by a simplified version of moment matching, Riggs (1989) suggested that the lower limit L and the upper limit U should be respectively estimated by the minimum and maximum observations in the random sample $\{X_i\}$; therefore only the shape parameters a and b must be estimated by matching the mean and variance of the fitted beta distribution [see (4) and (5)] to the computed mean and variance of the sample. This approach results in a linear system of two equations in the two unknowns a and b that can be solved easily; (c.f. Hahn and Shapiro, page 95, 1967). There are two fundamental problems with this approach. First, the fitted density vanishes at the largest and smallest observed values in the sample, which implies that these values would have negligible probability of occurring with the fitted input model. And second, substantial flexibility in the fitted distribution is lost since there is no capability to match the skewness and kurtosis observed in the sample. Thus more general moment-matching procedures for estimating beta distributions were considered.

Since the generalized beta distribution has $k = 4$ parameters, the first four theoretical moments specified by (4)–(7) must be matched to the corresponding sample moments computed from the data set $\{X_i\}$ as given as follows:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{8}$$

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{9}$$

$$\hat{\alpha}_3 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S}\right)^3 \tag{10}$$

$$\hat{\alpha}_4 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S}\right)^4 \tag{11}$$

The following system of nonlinear equations must then be solved to determine $\boldsymbol{\Theta}$:

$$\mu(a, b, L, U) = \mu(\boldsymbol{\Theta}) = \bar{X} \tag{12a}$$

$$\sigma^2(a, b, L, U) = \sigma^2(\boldsymbol{\Theta}) = S^2 \tag{12b}$$

$$\alpha_3(a, b) = \alpha_3(\boldsymbol{\Theta}) = \hat{\alpha}_3 \tag{13a}$$

$$\alpha_4(a, b) = \alpha_4(\boldsymbol{\Theta}) = \hat{\alpha}_4 \tag{13b}$$

To solve the nonlinear system of equations defined by (12) and (13) for the moment-matching estimates $\hat{L}$, $\hat{U}$, $\hat{a}$, and $\hat{b}$, a two-stage approach is conventionally used. In the first stage, system (13) is solved for $\hat{a}$ and $\hat{b}$ by iteratively minimizing the sum of squared differences between the left- and right-hand sides of (13); then resulting values of $\hat{a}$ and $\hat{b}$ are substituted into the linear subsystem (12) to solve for $\hat{L}$ and $\hat{U}$. In practice it was found that this procedure can yield infeasible estimates of the lower or upper limits such that

$$\min_i \{X_i\} < \hat{L} \qquad \text{or } \max_i \{X_i\} > \hat{U} \tag{14}$$

This infeasibility condition implies that the observed sample $\{X_i\}$ has zero probability of occurring under the fitted input model $f(x; \hat{\boldsymbol{\Theta}})$—and it is the writers' belief that such a logical inconsistency is completely unacceptable from both a theoretical and practical standpoint.

Although a modification of this conventional two-stage moment-matching procedure has been investigated and developed that was specially designed to guarantee feasibility of the final estimates $\hat{L}$ and $\hat{U}$ by eliminating the requirement to match the sample variance in (12), this modified procedure has also proved to be unreliable in practice. Specifically, both the conventional two-stage moment-matching procedure as well as the modified procedure to a suite of 80 sample data sets representing a broad spectrum of input processes that arise in the simulation of construction projects has been applied (AbouRizk 1990). In a substantial number of these data sets, the fitted beta distributions were clearly unacceptable even though excellent fits could be obtained by other methods. These results led to the development of an alternative implementation of moment matching that is designed for fitting a beta distribution to sample data by matching all four sample moments as closely as possible while avoiding the infeasibility condition (14).

Feasibility-Constrained Moment Matching

The following moment-matching (MM) technique for fitting a generalized beta distribution to sample data subject to a feasibility constraint on the estimated lower and upper limits of the fitted distribution is being proposed. A general nonlinear optimization algorithm must be used to solve the constrained minimization problem

$$\text{Minimize } [\mu(\boldsymbol{\Theta}) - \bar{X}]^2 + [\sigma^2(\boldsymbol{\Theta}) - S^2]^2 + [\alpha_3(a, b) - \hat{\alpha}_3]^2 + [\alpha_4(a, b) - \hat{\alpha}_4]^2 \tag{15}$$

$$\text{Subject to } a > 0 \tag{16a}$$

$$b > 0 \tag{16b}$$

$$L < \min\{X_i: i = 1, \ldots, n\} \tag{16c}$$

$$U > \max\{X_i : i = 1, \ldots, n\} \tag{16d}$$

The final solution (or approximate solution) to (15) and (16) yields the feasibility-constrained moment-matching estimates $\hat{a}$, $\hat{b}$, $\hat{L}$, $\hat{U}$. Although this approach is not guaranteed to yield an exact match to any of the sample moments, it has been found to yield generally superior fits compared to the other variants of moment matching described above. AbouRizk (1990) gives a comprehensive analysis of the performance of this feasibility-constrained moment-matching procedure in a suite of 80 data sets arising in the simulation of construction projects.

Maximum Likelihood

If the end points L and U of a given beta distribution are known and if $\{X_i: i = 1, 2, \ldots, n\}$ is a random sample of size n from that distribution, then the corresponding likelihood equations for the the shape parameters a and b are

$$\psi(a) - \psi(a + b) = \ln(G_1) \tag{17a}$$

$$\psi(a) - \psi(a + b) = \ln(G_2) \tag{17b}$$

where

$$G_1 = \left[\prod_{i=1}^{n} \left(\frac{X_i - L}{U - L}\right)\right]^{1/n} \tag{18}$$

$$G_2 = \left[\prod_{i=1}^{n} \left(\frac{U - X_i}{U - L}\right)\right]^{1/n} \tag{19}$$

and $\psi(z) = d/dz$ for $\ln[\Gamma(z)]$ is the digamma function [see (2)]. A number of methods could be used to solve (17). Johnson and Kotz (1970) suggested using a trial and error method. The Newton-Raphson method can also be used. BetaFit uses a technique developed by Beckman and Tietjen (1978) that does not require starting values for a and b.

The main drawback of the method of maximum likelihood (MLE) is that it requires prior knowledge of the end points of the beta distribution. In most construction engineering simulation applications, such knowledge is not usually available. In practice, the end points are fixed at some arbitrary values L and U such that the range $[L, U]$ contains all of the sample observations $\{X_i\}$. As noted previously, such an arbitrary approach frequently results in substantial loss of flexibility in the fitted beta distribution. Furthermore, the method of maximum likelihood is extremely sensitive to the values of L and U. Unless the simulation analyst has extensive knowledge of the properties of the underlying distribution so that the exact values of the end points can be verified, this method is not recommended.

Least-Squares Estimation of Beta CDF

Wilson (1983) proposed a regression-based method for fitting CDFs from the Johnson distribution system based on sample data. Swain et al. (1988) extended this estimation method and implemented it as part of an interactive software package for fitting Johnson distributions. In this subsection an analogous method for fitting beta distributions based on sample data is considered.

If the beta CDF in (3) accurately represents the underlying distribution

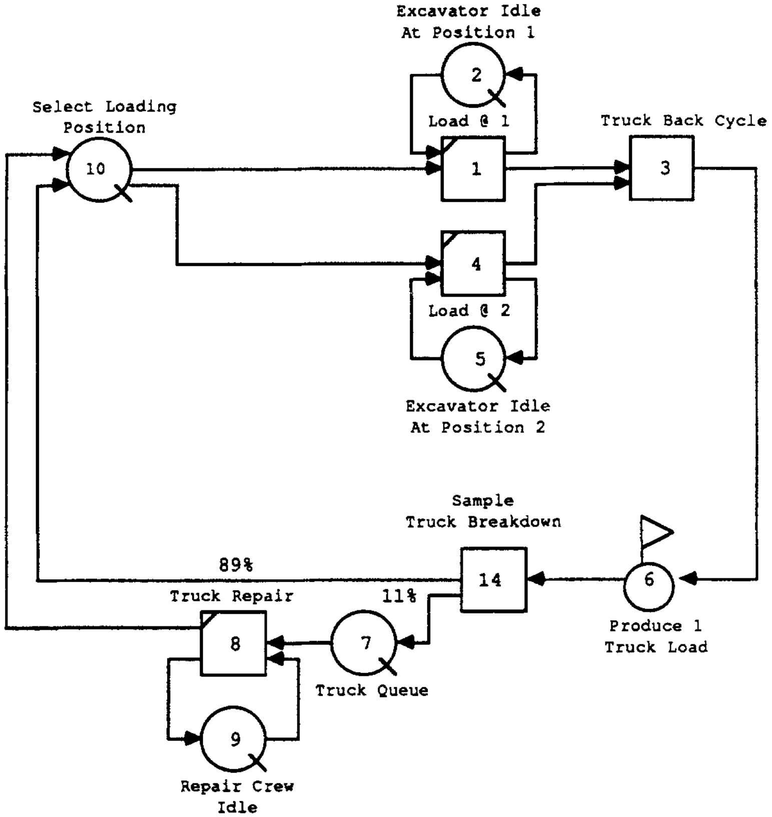


FIG. 1. CYCLONE Model of Earthmoving Operation

of the random sample $\{X_i: i = 1, \ldots, n\}$ and if the corresponding order statistics are denoted $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, the following nonlinear statistical model for the jth order statistic in the sample is obtained:

$$F[X_{(j)}; \boldsymbol{\Theta}] = \frac{j}{(n+1)} + \varepsilon_j \qquad \text{for } j = 1, \ldots, n \qquad (20)$$

where each "error" ε_j has mean $E(\varepsilon_j) = 0$. The basic idea is to view (20) as a nonlinear regression model—even though the unknown parameter vector $\boldsymbol{\Theta}$ appears on the left-hand side as part of the dependent variable $F[X_{(j)}; \boldsymbol{\Theta}]$ rather than on the right-hand side with the independent variable $j/(n + 1)$. Although this setup appears to be unconventional, it allows for a legitimate application of the principle of least-squares estimation.

The covariance between jth and kth "errors" in (20) is

$$\mathrm{cov}(\varepsilon_j, \varepsilon_k) = \Delta_{jk} = \frac{j(n-k+1)}{(n+1)^2(n+2)} \qquad \text{for } 1 \leq j \leq k \leq n \qquad (21)$$

and this suggests using weighted nonlinear regression to estimate the unknown parameter vector $\boldsymbol{\Theta}$ by minimizing the sum of squares

$$\sum_{j=1}^{n} W_j \left\{ F[X_{(j)}; \boldsymbol{\Theta}] - \frac{j}{n+1} \right\}^2 \qquad (22)$$

TABLE 1. Sample Observations and Statistics for Dozer and Truck Cycle Times

Observation (1)	Dozer cycle (min) (2)	Truck cycle (min) (3)
1	1.55	10.2
2	1.53	9
3	1.20	8
4	1.16	7.9
5	0.90	10.5
6	1.15	11
7	0.90	13
8	1.11	14
9	1.10	15
10	0.88	11
11	1.10	10.6
12	0.87	8.9
13	1.09	7.9
14	1.25	10.2
15	1.06	11.5
16	1.05	12.9
17	1.20	13.5
18	1.03	11.4
19	1.01	9.8
20	1.00	7.8
21	1.62	8
22	1.50	9.4
23	1.00	11.9
24	1.30	8.2
25	0.27	12.9
26	0.52	11.9
27	0.48	10.3
28	0.30	11.6
29	1.33	8.2
30	1.00	9.8
31	1.40	9.5
32	1.00	7.6
33	1.38	11.5
34	0.95	10.4
35	1.20	10.9
36	0.89	10
37	1.45	10
38	0.86	11.9
39	1.60	10.4
40	—[a]	10.5
41	0.30	10.9
42	0.25	11
43	0.85	12
44	0.85	14
45	0.19	13
46	1.36	8.9
47	1.40	7.6
48	1.00	8.2
49	0.83	7.6

TABLE 1. (*Continued*)

(1)	(2)	(3)
50	0.83	7.8
51	1.90	8
52	1.80	9.9
53	0.82	10.4
54	0.67	10.7
55	0.58	10.8
56	0.82	11.3
57	0.66	9.6
58	0.49	7.9
59	0.67	9.5
60	0.80	7.8
61	0.59	7.8
62	0.26	7.9
63	1.37	10.2
64	0.55	9.5
65	0.59	7.9
66	0.75	8.9
67	0.70	8.9
68	0.70	6.9
69	0.65	10.4
70	0.64	10.8
71	0.62	12
72	0.60	9.6
73	0.55	8.7
74	—	8.9
75	—	10.5
76	—	14.5
77	—	8.7

[a] Data missing.

TABLE 2. Beta Parameter Estimates from Various Methods for Dozer Cycle Time

Parameter (1)	Maximum likelihood (2)	Moment matching (3)	Ordinary least squares (4)	Diagonally weighted least squares (5)
Low	0.1843	0.0612	0.1781	0.1793
High	1.9570	2.0621	1.9372	1.9389
a	1.5385	4.4993	1.7631	1.7680
b	2.0970	6.4053	2.2502	2.2466
K-S[a]	0.0856	0.1230	0.0633	0.0663
At X	0.55	1.0	1.11	1.11

[a] K-S = Kolmogrov-Smirnov statistic.

for some choice of the weights $\{W_j: j = 1, \ldots, n\}$. In this work, two variants of the method of least-squares estimation were applied. (a) The ordinary least squares (OLS) procedure uses constant weights so that $W_j = 1$ for $j = 1, \ldots, n$. (b) The diagonally weighted least squares (DWLS) procedure uses weights that are inversely proportional to the corresponding "error" variances so that $W_j = 1/\text{var}(\varepsilon_j) = 1/\Delta_{jj}$ for $j = 1, \ldots, n$.

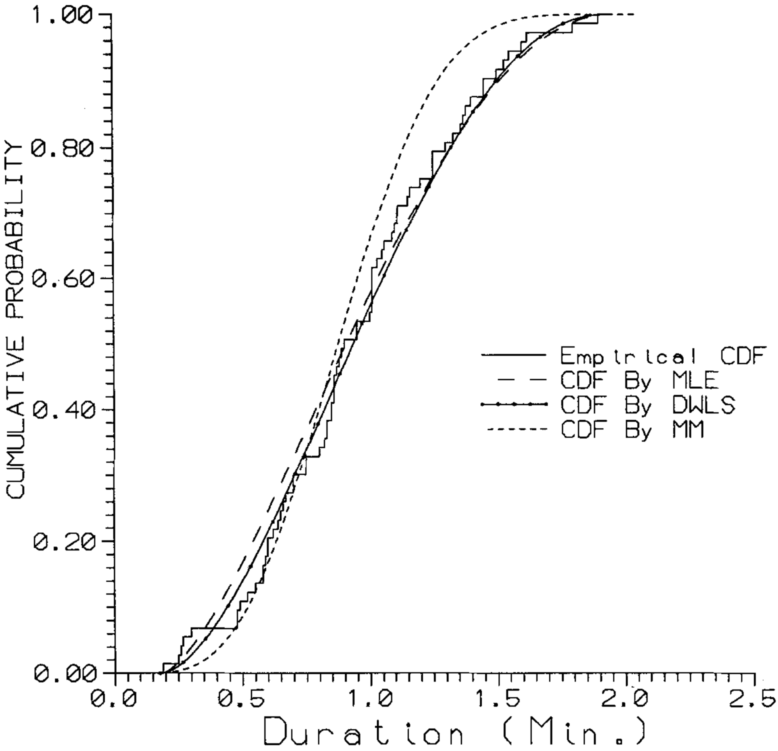


FIG. 2. Fitted Beta Distributions by MLE, MM, and DWLS to Dozer Cycle Times

The OLS estimation procedure uses a general nonlinear optimization algorithm to solve the constrained minimization problem

$$\text{Minimize} \sum_{j=1}^{n} \left\{ F[X_{(j)}; \Theta] - \frac{j}{n+1} \right\}^2 \tag{23}$$

$$\text{Subject to } a > 0 \tag{24a}$$

$$b > 0 \tag{24b}$$

$$L < X_{(1)} = \min\{X_i\} \tag{24c}$$

$$U > X_{(n)} = \max\{X_i\} \tag{24d}$$

Similarly, the DWLS estimation procedure uses a general nonlinear optimization algorithm to solve the constrained minimization problem

$$\text{Minimize} \sum_{j=1}^{n} \frac{\left\{ F[X_{(j)}; \Theta] - \frac{j}{n+1} \right\}^2}{\Delta_{jj}} \tag{25}$$

subject to (24).

Note that both of these techniques require a starting point for Θ, and in general this starting point must be computed by some other method.

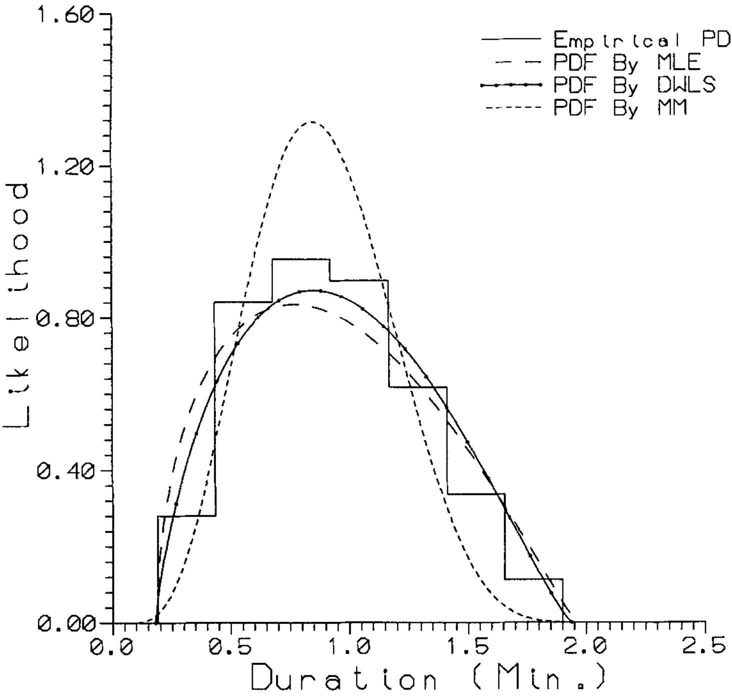


FIG. 3. PDF of Beta Distributions Fitted by MLE, MM, and DWLS to Dozer Cycle Times

For fitting beta distributions based on sample data, the OLS and DWLS estimation procedures possess several distinct advantages. First, prior knowledge of the end points L and U is not required, but such information can be easily incorporated into the estimation procedure if it is available. Second, when the end points are unknown, the feasibility of the final estimates $\hat{L}$ and $\hat{U}$ is guaranteed by the last two constraints in (24). Third, the behavior of the final estimated parameter vector $\hat{\Theta} = (\hat{L}, \hat{U}, \hat{a}, \hat{b})$ does not depend critically on the starting point used in the minimization algorithm for (24) or (25). The principal disadvantage of the OLS and DWLS estimation procedures is that they can require substantially more computation time than moment matching or maximum likelihood, especially in large data sets.

EXAMPLE APPLICATION

To illustrate how the discussed beta distribution fitting methods can be used, an example from the field of construction engineering is considered. BetaFit (described in Appendixes I and II) contains the programming implementations of the techniques described in the present paper and therefore will be used in this example. A construction engineer conducts a simulation experiment for equipment allocation (e.g. trucks and dozers) on a heavy-construction operation. A typical CYCLONE simulation model for such an operation is shown in Fig. 1. The MicroCYCLONE simulation system (Halpin 1990) can be used to carry the simulation. Due to various random factors

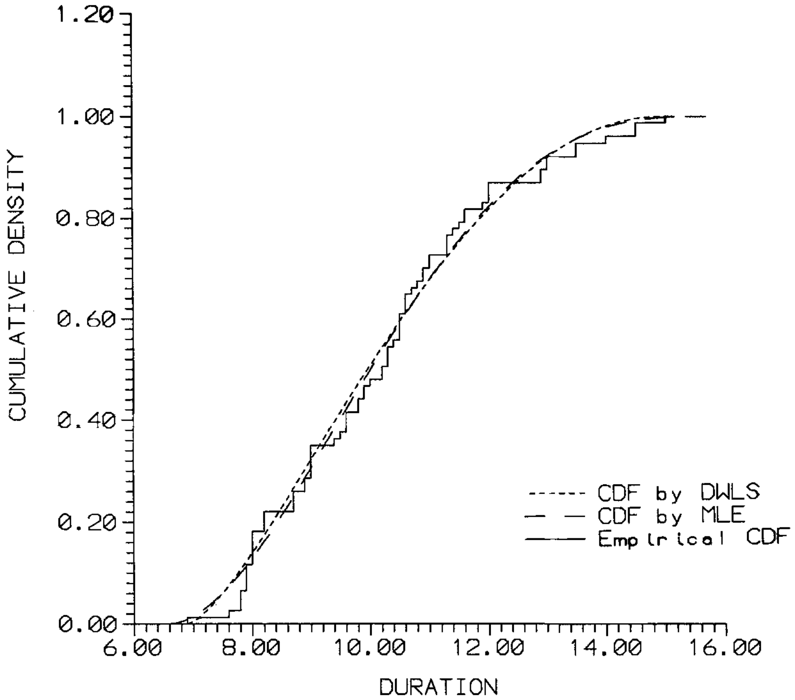


FIG. 4. Fitted Beta Distributions by MLE and DWLS to Truck Cycle Times

involved in the process, the engineer decides to model the cycle time of the dozer operation and the truck back cycle time as a random process. Data is collected for the ongoing operation using a stop-watch study or by reviewing recorded film for short cycle operations (e.g. dozing). The tabulated cycle times given in Table 1 for the dozer and truck cycles were obtained from Al-Masri (1985). The origin and specific data collection mechanisms of the tabulated cycle times can be found in Al-Masri (1985). The material that was hauled was pre-blasted and consisted of common earth or rock. It should be noted that this example illustration does not elaborate on filtering the data after its collection to ensure a close representation of the system being modeled in the simulation experiment is attained. In general, the collected data will be sorted and carefully examined. Unrealistic values (outliers) are properly dealt with (e.g. eliminated). In addition, the scope of the activity being modeled should be carefully defined (i.e. start and end of the cycle times) to ensure that the quality of the data obtained is within acceptable limits. If the collected sample is of inferior quality containing many nonrepresentative cycle times the input modeling method used would have very little significance as the model will normally be inaccurate regardless of the fitting procedure employed. As such the accuracy of the fitting method used cannot by itself secure appropriate simulation results of the system being modeled.

In this sample application it is assumed that the data is of the required value and that the cycle time is representative of the actual activity being modeled. The engineer decides to use a beta distribution to model the cycle-time for suitability of beta distribution to model such random processes in

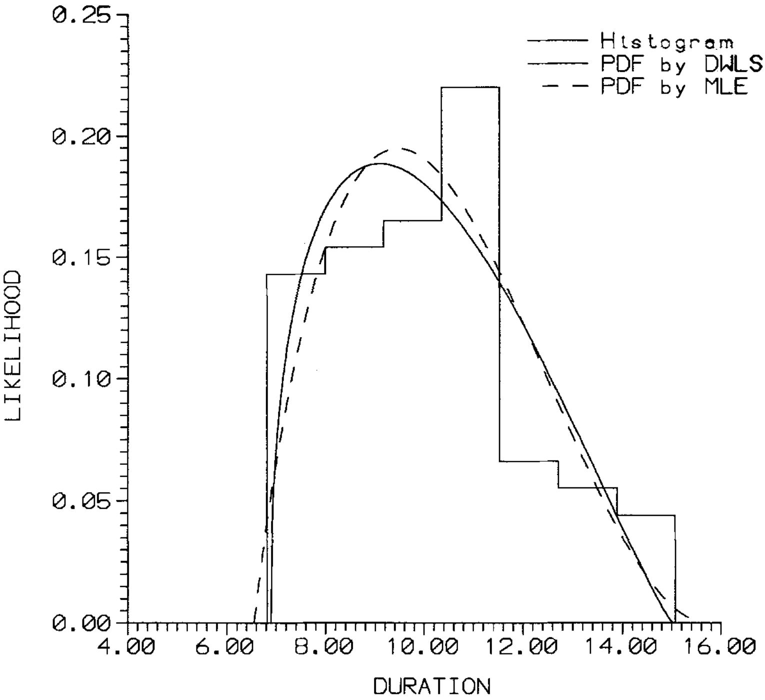
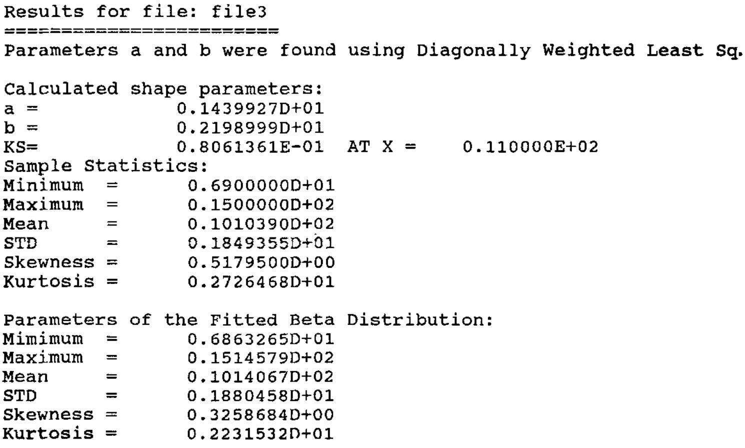


FIG. 5. PDF of Beta Distributions Fitted by MLE and DWLS for Truck Cycle Times



```
Results for file: file3
=======================
Parameters a and b were found using Diagonally Weighted Least Sq.

Calculated shape parameters:
a =          0.1439927D+01
b =          0.2198999D+01
KS=          0.8061361E-01  AT X =    0.110000E+02
Sample Statistics:
Minimum  =     0.6900000D+01
Maximum  =     0.1500000D+02
Mean     =     0.1010390D+02
STD      =     0.1849355D+01
Skewness =     0.5179500D+00
Kurtosis =     0.2726468D+01

Parameters of the Fitted Beta Distribution:
Mimimum  =     0.6863265D+01
Maximum  =     0.1514579D+02
Mean     =     0.1014067D+02
STD      =     0.1880458D+01
Skewness =     0.3258684D+00
Kurtosis =     0.2231532D+01
```

FIG. 6. Analysis Results of Truck Cycle Times as Outputted from BetaFit

construction. BetaFit was used to fit a beta distribution to the collected sample of observations.

Analysis of the dozer cycle time using the four different procedures available in BetaFit yielded different parameter estimates as expected. To facilitate comparison between the quality of the fits attained by the various methods the DWLS plot was used as a base since it resulted in the least

Kolmogrov-Smirnov (KS) statistic as shown in Table 2. The empirical CDF and fitted CDFs by DWLS, MLE, and MM are shown in Fig. 2. The corresponding PDFs are given in Fig. 3. The OLS plot was very close to the DWLS plot and, therefore, was not shown in the figure to avoid cluttering. A close visual assessment shows that the DWLS method provides the closest tracking of the empirical CDF. The results of the KS test given in Table 2 show that the maximum "gap" was about 0.06 and occurred at $X = 1.11$ minutes with both least squares procedures, whereas it was 0.08 at $X = 0.55$ min and 0.12 at $X = 1.0$ min for the MLE and moment matching procedures, respectively. The fit provided by the least squares methods was found to be the best of the four fitted distributions both visually and because it resulted in the least KS values and, therefore, the engineer decides to use the parameters of the beta distribution corresponding to DWLS [i.e. Beta (0.18, 1.94, 1.77, 2.25)] in the MicroCYCLONE simulation experiment.

The analysis of the truck cycle was carried in a similar way with the resulting CDF and PDF plots for the MLE and DWLS approaches given in Figs. 4 and 5, respectively. The results of the OLS procedure were very close to the DWLS and therefore not included in the graph whereas the MM approach yielded an unacceptable plot and was eliminated. The distribution parameters from either the MLE or DWLS can be used in the simulation study as both yielded acceptable fits. The results of the DWLS analysis as outputted from BetaFit is given in Fig. 6 for illustration. The parameters of the resulting beta distribution are [Beta (6.86, 15.15, 1.44, 2.20)]. This may now be used in the simulation experiment to specify the truck cycle time.

GENERAL CONCLUSIONS

This paper presented numerical techniques that can be used to fit beta distributions to sample data for construction engineering and statistical applications. The methods were implemented in BetaFit which provides an easy, accurate and efficient way to fit beta distribution for various applications. In general terms, the least-square procedures presented herein consistently yielded equally good or better fits compared to MLE and MM. Although a formal Monte Carlo study was not performed to numerically evaluate the various procedures, the methods were applied to 80 different construction data sets. Visual assessment of the fits showed that in most of the cases, the least-squares procedures (with function evaluation at every 5th point) were very competitive with the fits obtained from MLE (with end points assumed shifted 5%). In very few cases was the fit better with MLE. Moment matching in both versions did not show any advantage except in terms of convergence speed compared to the least squares procedures. Unlike MLE, the least-squares procedures do not require preknowledge of the end points of the distribution. This proves to be very advantageous in applications where such knowledge is not immediately available (e.g. construction-duration data).

The importance of input modeling cannot be overemphasized. Simulation results with inappropriate input models are suspect and should be dealt with carefully. The techniques presented in this paper provide the simulation analyst with an easy-to-use tool that provides robust and accurate input models to a simulation experiment. The implementation in BetaFit provides an easy-to-use tool that requires minimal effort to produce sound models of random processes.

APPENDIX I. GENERAL OVERVIEW OF BETAFIT

BetaFit is the programming implementation of the fitting techniques discussed in this paper. The operations of BetaFit can be summarized as follows:

- BetaFit reads a set of data from a sequential ASCII file.
- The statistics of the sample are computed.
- A beta distribution is then fitted to the sample based on the user's choice of the fitting procedure.
- A report of the session is produced and various plot files are generated for the fitted and empirical distributions.

BetaFit Operations

An input file containing the data to be analyzed should be created by the user. After entering the input file name, the program reads the data points from the specified input file sequentially. The observations are then internally sorted, and the sample statistics calculated. One of the available techniques should be chosen based on the user's preference. The program will perform the required computations and fits the best possible beta distribution using the selected technique. Plot-files are then generated. Files currently supported are: frequency histogram constructed from the sample, empirical CDF, fitted beta PDF, and fitted beta CDF.

At the end of the session, BetaFit generates a report that is printed to the screen and to an ASCII file. The report includes: the name of the file, the method employed in estimating the parameters of the fitted distribution, the sample statistics, the parameters of the fitted beta distribution, and the Kolmogrov-Smirnov statistic as shown in Fig. 6.

APPENDIX II. PROGRAMMING IMPLEMENTATIONS

A brief description of the important programming implementations are discussed in this appendix. A more detailed discussion is available in AbouRizk (1990).

Preparation

Data should be stored in a plain ASCII file. The delimiters can be a "," or a carriage return mark.

Input

Name of the file containing data. Fitting method to be used.

Preliminary Computations

Statistics of the sample are computed using (8)–(11). Data is sorted in ascending order. Starting values of a and b are evaluated. This is required for function minimization used in solving systems (12), (15), (24), and (25). The following is used: fix the end points of the distribution to the lower and upper sample points and match the mean and variance of the beta distribution to those of the sample. Solve system (12) yielding

$$\hat{a}_1 = \frac{\mu - L}{U - L}\left[\frac{(\mu - L)(U - \mu)}{\sigma^2} - 1\right] \tag{26}$$

$$\hat{b}_1 = \hat{a}_1 \frac{U - \mu}{\mu - L} \tag{27}$$

Solve for L, U, a, and b, depending on the method chosen. If the case is MM, minimize (15) subject to (16) using the Nelder-Mead method described later herein. If the case is MLE, use the method described by Beckmen and Tietjen (1978) with modifications suggested in (Griffiths and Hill 1985). If the case is OLS, minimize (24) subject to (25) using the Nelder-Mead method. The nonintegratable cumulative density of the beta distribution $F(X_{(j)}; \hat{\Theta}_i)$ is evaluated as described by Majumder and Bhattacharjee (see Griffiths and Hill 1985). If the case is DWLS, a similar approach to OLS is used except for the function to be minimized.

Nelder-Mead Method

The procedure is a direct search method which evaluates a function of n variables at $n + 1$ vertices of a general simplex. The simplex is then moved away from large values, extended or contracted depending on the contours of the response surface (Olsson 1974). The implementation is based on the Nelder-Mead procedure described by Olsson (1974). This iterative minimization procedure is repeated until one of the following conditions are met: (1) The function reaches the minimum value 10^{-6}; (2) the added enhancement values of the parameters $\hat{\Theta}$ is less than 10^{-35}; and (3) the total number of iterations exceed 400.

Evaluating Beta PDF

The PDF of the beta distribution as given in (1) requires estimation of the gamma function given by (2). The method of Pike and Hill (1984) is used to evaluate the gamma function. Furthermore, the natural logarithm of the complete beta function is evaluated rather than the function itself as recommended by Cran et al. (Griffiths and Hill 1985) as follows:

$$\ln(\beta) = \ln \Gamma(a) + \ln \Gamma(b) - \ln \Gamma(a + b) \tag{28}$$

where β = complete beta function; and Γ = gamma function.

APPENDIX III. REFERENCES

AbouRizk, S. M. (1990). "Input modeling for construction simulation," PhD thesis, Purdue Univ., West Lafayette, Ind.

AbouRizk, S., and Halpin, D. (1992a). "Statistical properties of construction duration data." *J. Constr. Engrg. and Mgmt.*, ASCE, 118(3), 525–544.

AbouRizk, S., and Halpin, D. (1992b). "Modeling input data for construction simulation." *Proc., 8th Annu. Conf. on Comp. in Civ. Engrg.*, ASCE, New York, N.Y., 1147–1154.

AbouRizk, S., Halpin, D., and Wilson, J. (1992). "Visual interactive estimation of beta distributions." *J. Constr. Engrg. and Mgmt.*, ASCE, 117(4), 589–605.

Al-Masri, F. M. (1985). *Analysis of a loader-truck system operation using the computer simulation language SIMAN.* Pennsylvania State Univ., University Park, Pa.

Beckman, R. J., and Tietjen, G. L. (1978). "Maximum likelihood estimation for the beta distribution." *J. Statist. Comput. Simul.*, 7, 253–258.

Farid, F., and Aziz, T. (1993). "Simulating paving fleets with non-stationary travel." *5th Int. Conf. on Comp. in Civ. and Build. Engrg.*, 1198–1206.

Griffiths, P., and Hill, I. D. (1985). *Applied statistics algorithms.* John Wiley and Sons, New York, N.Y.

Hahn, G. J., and Shapiro, S. S. (1967). *Statistical models in engineering.* John Wiley and Sons, New York, N.Y.

Halpin, D. W. (1990). *MicorCYCLONE user's manual.* Division of Construction Engineering and Management, Purdue Univ., West Lafayette, Ind.

Johnson, N. L. (1948). "System of frequency curves generated by methods of translation." *Biometrica*, 36, 149–176.

Johnson, N. L., and Kotz, S. (1970). *Continuous univariate distributions.* John Wiley and Sons, New York, N.Y.

Olsson, D. M. (1974). "A sequential simplex program for solving minimization problems." *J. Quality Technol.*, 6, 53–57.

Pike, M. C., and Hill, I. D. (1964). "Algorithm 291." *Comm. ACM*, 9, 684.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., and Mykytka, E. F. (1979). "A probability distribution and its uses in fitting data." *Technometrics*, 21, 201–214.

Riggs, L. S. (1989). "Numerical approach for generating beta random variables." *J. Comp. in Civ. Engrg.*, ASCE, 3(2), 183–191.

Schmeiser, B. W., and Deutsh, S. J. (1976). "A versatile four parameter family of probability distributions suitable for simulation." *AIIE Trans.*, 9(2), 176–181.

Swain, J., Venkatraman, S., and Wilson, J. (1988). "Least squares estimation of distribution functions in Johnson's translation system." *J. Statist. Comput. Simul.*, 29, 271–297.

Touran, A., and Wiser, E. (1992). "Monte Carlo technique with correlated random variables." *J. Constr. Engrg. and Mgmt.*, ASCE, 118(2), 258–272.

Tukey, J. W. (1960). "The practical relationship between the common transformations of percentages of counts and of amounts." *Tech. Rep. No. 36.* Statistical Techniques Research Group, Princeton University, Princeton, N.J.

Wilson, J. (1983). "Modeling multivariate populations with Johnson translation systems." *Tech. Rep.*, Mechanical Engineering Department, University of Texas at Austin, Austin, Tex.

APPENDIX IV. NOTATION

The following symbols are used in this paper:

a = first shape parameter of beta distribution;
b = second shape parameter of beta distribution;
$E(t)$ = expected value of variable t;
$F(x, \boldsymbol{\Theta})$ = cumulative beta distribution function evaluated at cutoff value x;
$f(x, \boldsymbol{\Theta})$ = beta probability density function evaluated at the cutoff value x as defined in Eq. (1);
G_1, G_2 = constants evaluated from sample of observations as defined in Eqs. (18) and (19);
L = lower end point of beta distribution;
m = mode of beta distribution;
n = total number of observation in sample;
S^2 = variance of sample of observations as defined in Eq. (9);
U = upper end point of beta distribution;
W_j = weight factor applied to nonlinear model as defined in Eq. (22);
$\bar{X}$ = mean of sample of observations as defined in Eq. (8);
α_3 = coefficient of skewness of given population;
α_4 = kurtosis of given population;
$\Gamma(z)$ = gamma function of variable z as defined in Eq. (2);
Δ_{jk} = the covariance between jth and kth "errors" as defined in Eq. (21);
ε = error term in regression model as defined in Eq. (20);
$\boldsymbol{\Theta}$ = vector of parameters defining generalized beta density;

μ = mean of given population;
σ = standard deviation of given population;
σ^2 = variance of given population; and
$\psi(t)$ = digamma function of the variable t, i.e. derivative of logarithm of gamma function.

Subscripts

i = index counting number of observations in sample; and
j = index counting ordered sample of observations.

Overscores

^ = indicator for subjective estimate (like $\hat{L}$) of property (like L) of given random variable.