Proceedings of the 1996 Winter Simulation Conference
ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain

EXPERTFIT: TOTAL SUPPORT FOR SIMULATION INPUT MODELING

Averill M. Law
Michael G. McComas

Averill M. Law and Associates
P.O. Box 40996
Tucson, Arizona 85717

ABSTRACT

In this paper, we explain the important role of simulation input modeling in a successful simulation study. Two pitfalls in simulation input modeling are then presented and we explain how any analyst, regardless of their knowledge of statistics, can easily avoid these pitfalls through the use of ExpertFit, the Windows-based successor to the UniFit II input modeling package. We use a set of real-world system data to demonstrate how the package automatically specifies, evaluates, and ranks candidate probability distributions, and then assists an analyst in deciding whether the "best" candidate probability distribution provides an adequate representation of the data. If no candidate probability distribution provides an adequate fit, then ExpertFit can define an empirical distribution function. In either case, the probability distribution can be automatically expressed in the analyst's simulation software. We then consider the general case of selecting a probability distribution in the absence of data. As an example, we show how ExpertFit can be used to create busy-time and downtime models for machines that are subject to random breakdowns.

1 THE ROLE OF SIMULATION INPUT MODELING IN A SUCCESSFUL SIMULATION STUDY

In this section we will describe simulation input modeling and show consequences that might result if this important, but sometimes neglected, activity is performed improperly. We then suggest that with the use of ExpertFit any simulation analyst can perform simulation input modeling more quickly and with greater accuracy than would otherwise be possible.

1.1 The Nature of Simulation Input Modeling

One of the most important activities in a successful simulation study is that of representing each source of system randomness by a probability distribution. For example in a manufacturing system, processing times, operating times before failure, and repair times for a machine should usually be modeled by probability distributions.

In this paper, we use the phrase "simulation input modeling" to mean the process of choosing a probability distribution for each random component of the system under study and expressing this representation in a form that can be used with the analyst's choice of simulation software. In Sections 2 and 3 we will demonstrate how an analyst can easily and accurately choose an appropriate probabilistic representation using the ExpertFit software.

1.2 Two Pitfalls in Simulation Input Modeling

The authors have identified a number of pitfalls that can undermine the success of simulation studies (Law, McComas, and Vincent 1994, Law and Kelton 1991, Law and McComas 1989). Two of these pitfalls relate directly to simulation input modeling and are summarized in this section.

1.2.1 Pitfall Number 1: Replacing a Distribution by its Mean

Simulation analysts have sometimes replaced an input probability distribution by its mean in their simulation models. This practice may be caused by a lack of understanding on the part of the analyst or by lack of information on the actual form of the distribution (e.g., only an estimate of the mean of the distribution is available). Such a practice may produce completely erroneous results, as is shown by the following example.

Consider a manufacturing system consisting of a single machine tool at which jobs arrive to be processed. Suppose that the mean interarrival time of jobs is one minute and the mean processing time is 0.99 minute. Suppose further that the interarrival times and processing times actually have an exponential distribution. Then it can be shown that the long-run mean number of jobs waiting in the queue is *approximately 98*. On the other hand, suppose we were to follow the dangerous practice of replacing a source of randomness with a constant value. If we assume that each interarrival time is *exactly* one minute and each processing time is *exactly* 0.99 minute, *then each job is finished before the next arrives and no job ever waits in the queue*! The variability of the probability distributions, rather than just their means, has a significant impact on the congestion level in most queueing-type (e.g., manufacturing) systems. In Section 2 we shall show how the use of ExpertFit makes choosing an appropriate probability distribution a simple and easy process.

1.2.2 Pitfall Number 2: Incorrect Modeling of Random Machine Downtimes

The largest source of randomness for many manufacturing systems is that associated wth random machine downtimes. An analyst is often faced with representing in a simulation model the random machine downtimes of a machine that has not yet been purchased. Data concerning the actual downtime behavior of machine tools is, thus, unavailable and the analyst must rely on estimates of reliability provided by vendors and engineers. Suppose, for example, that a vendor claims that a machine tool will be down 10 percent of the time, but is unwilling or unable to provide more information on its operating time before breakdown and its repair time. Given the limited available information, some simulation analysts account for downtimes by simply reducing the machine processing rate by 10 percent. Law and McComas (1989) compare this practice to a more accurate model that we describe in Section 3. Although the two modeling approaches led to similar results for an average throughput measure of performance, the use of the reduced-production-rate model led to large errors with regard to measures such as average time in system and maximum number of jobs in queue. Accurate estimation of the latter performance measures is essential in many simulation studies. Thus, serious errors can result if an incorrrect, simplified approach is taken. We will show in Section 3 how easy it is to obtain a more accurate model of random machine downtimes using ExpertFit.

1.3 Advantages of Using ExpertFit

With the assistance of ExpertFit *any* analyst, regardless of their prior knowledge of statistics, can avoid the two pitfalls introduced above. When system data are available, a complete analysis with the package takes just minutes. The package identifies the "best" of the candidate probability distributions, and assists the analyst in deciding whether the fit is good. If none of the candidate distributions provides an adequate fit, then an empirical distribution function can be created by ExpertFit. In either case, the representation of system randomness can be automatically expressed in the analyst's choice of simulation software. Appropriate probability distributions can also be selected when no system data are available. For the important case of machine breakdowns, ExpertFit will determine appropriate busy-time and downtime probability distributions that match the system's behavior, *even if the machine is subject to blocking or starving.*

2 USING ExpertFit WHEN SYSTEM DATA ARE AVAILABLE

We now consider the case where an analyst has system data corresponding to the source of randomness to be represented in the simulation model. Our intention is to highlight the capabilities of ExpertFit. A demo disk with detailed commentary on program operation is available at no charge from the authors.

Three types of analyses are available for selecting probability models. In addition to the analysis of system data, there are two analysis types available when no system data are available (see Section 3). We have designed ExpertFit to embody our years of experience in selecting appropriate simulation input models. The user interface features multi-tabbed folders that correspond to the recommended steps in an analysis. Each tab organizes the appropriate options in a way that reflects our recommended analysis approach. Each option has default configuration settings that make it easy for an analyst to do any statistical procedure. All graphs are designed to assist in meaningful comparisons and to minimize possible analyst misinterpretation. For example: a) multiple models can be plotted on the same graph, b) error graphs are automatically scaled so that the visual size of an error reflects the severity of the error, and c) whenever possible, error bounds (safety limits) are displayed. These software features make it easy for any analyst to perform accurate and thorough analyses of data sets, regardless of their prior knowledge of statistics. On the other hand, the user interface is completely flexible so that an experienced analyst can easily access the full set of available tools for performing comprehensive and complete data set analyses in any order desired.

A data analysis is done using a folder with four tabs. The first tab has options for obtaining and displaying the features of a data sample; an analyst can read a data file, manually enter or edit a data set, paste in a data set from the system clipboard, as well as perform a variety of transformations. Once a sample is available, an analyst can create a number of graphical and tabular sample summaries, including histograms and plots designed to assess the randomness of the observations.

The data set we have chosen for this example consists of part processing times provided to us by a major automobile manufacturer.

At the second tab models can be fit to the sample. For the recommended guided fitting option, the basic information required by ExpertFit to begin the fitting and evaluation process is a specification of the range of the underlying random variable. For many data sets like the example processing times, the underlying random variable can be characterized as being greater than zero with no definite upper bound. ExpertFit responded to our choices by fitting distributions with ranges starting at zero and distributions whose lower endpoint was estimated from the data itself. These candidate models were then automatically evaluated. After a few seconds the result screen shown in Figure 1 was displayed.

ExpertFit fit and ranked 26 candidate models, with the three best-fitting models listed on the screen along with their scores. The displayed scores are calculated by a proprietary evaluation scheme that is based on our 18 years of research in this area, including the analysis of 35,000 computer-generated data sets. Results from the heuristics that we have found to be the best indicators of a good model fit are combined and the resulting numerical evaluation is normalized so that 100 indicates the best possible model and 0 indicates the worst possible model. These scores are *comparative* in nature and do not give an overall assessment of the quality of fit. ExpertFit provides a separate *absolute* evaluation of the adequacy of fit provided by the best-ranked model. This

Relative Evaluation of Candidate Models

Model	Relative Score	Model Range
1 - Inverted Weibull	100.00	Larger than 0
2 - Gamma(E)	92.00	Larger than 24.79809
3 - Log-Logistic(E)	90.00	Larger than 24.79809

26 models are defined with scores between 0.00 and 100.00

Absolute Evaluation of Model 1 - Inverted Weibull

Based on a heuristic evaluation, there is no current evidence for not using the primary model. If you are doing simulation, then the primary model will probably provide a good representation for your data. However, we recommend further confirmation of the primary model.

Additional Information Concerning Model 1 - Inverted Weibull

Result of an Anderson-Darling goodness -of-fit test at level .1	Do not reject
"Error" in the model mean relative to the sample average	-.09670 = .26%

Figure 1: Evaluation of Candidate Models for the Processing-Time Data

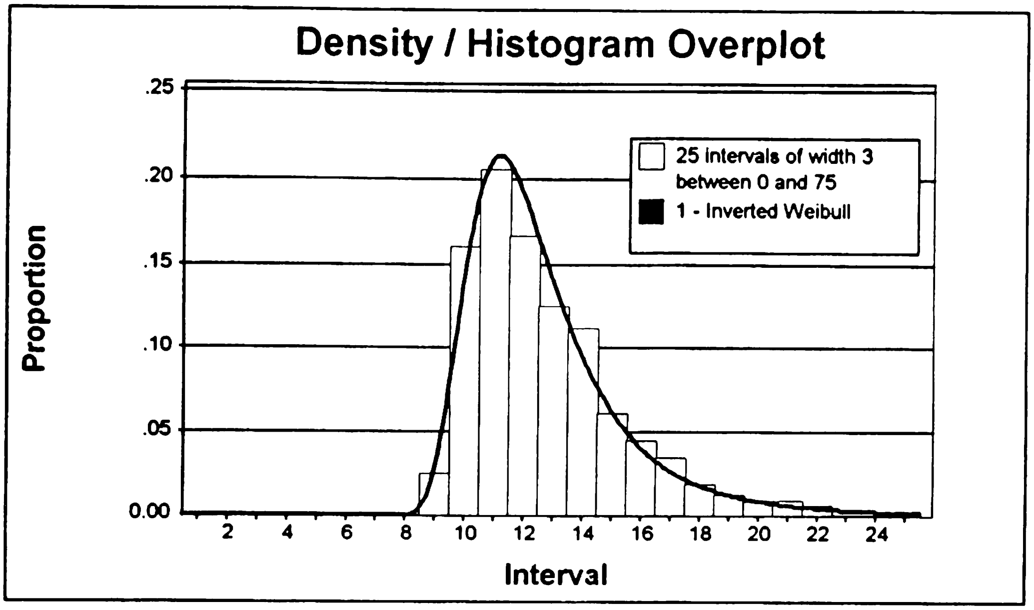

Figure 2: Density/Histogram OverPlot for the Processing-Time Data

absolute evaluation is critical because, perhaps, one third of all data sets are not well represented by a standard distribution. *Furthermore, ExpertFit is the only software package that provides such an absolute evaluation.*

In Figure 1 we see that the inverted Weibull distribution (range starts at zero) is the best model for the processing-time data. Although the inverted Weibull distribution may be unfamiliar to you, it is can be used in most simulation packages since it can be generated as the inverse of a Weibull random variable. It should also be noted that ExpertFit completed the entire analysis without further input from the analyst; only the range had to be specified.

After guided fitting, an analyst is automatically transferred to the third tab at which specified models can be compared to the sample to assess the quality of fit. Among our favorite comparisons are the density/histogram overplot and the distribution function differences plot, which are shown in Figures 2 and 3, respectively. In the former case, the density function of

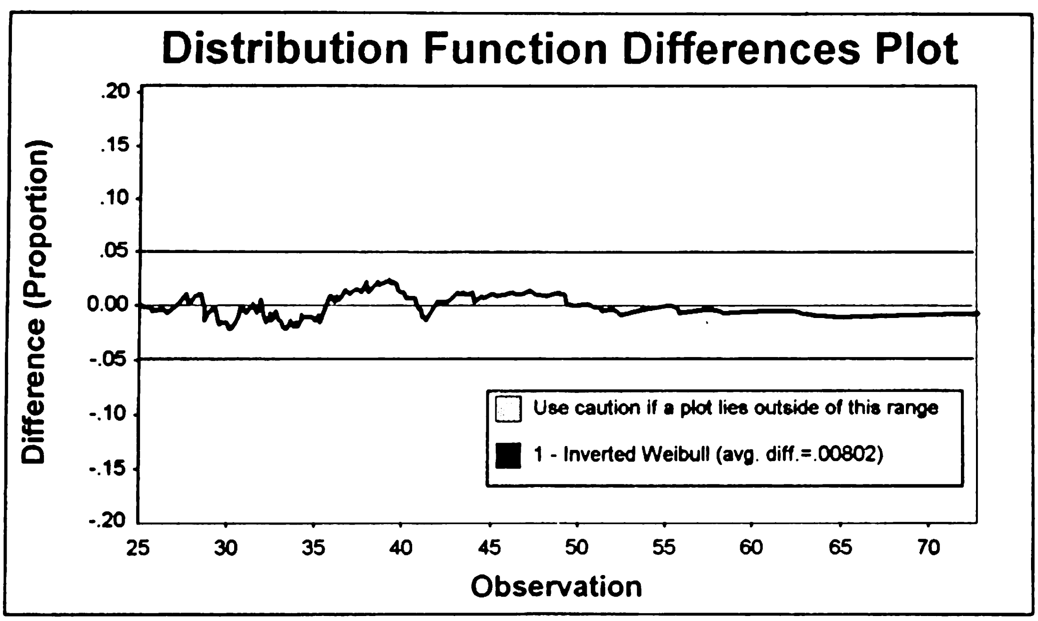

Figure 3: Distribution Function Differences Plot for the Processing-Time Data

Simulation Software	Representation
GPSS/H 3	RVIWEIB(<stream>,6.272056, 32.834140)
ProModel	InvWeibull(6.272056, 32.834140, <stream>, 0.000000)
Taylor II	1./weibull(0.028324, 6.272056)
WITNESS	1./WEIBULL(6.272056, 0.030456, <stream>)

Figure 4: Simulation Software Representation of the Inverted Weibull Distribution

the inverted Weibull distribution has been plotted over a histogram of the data (a graphical estimate of the true density function). This plot indicates that the inverted Weibull distribution is a good model for the observed data. The distribution function differences plot graphs the differences between a sample distribution function (a graphical estimate of the true distribution function) and the distribution function for the inverted Weibull distribution. Since these vertical differences are small (i.e., within the horizontal error bounds), this also suggests that the inverted Weibull distribution is a good representation for the data. Note that tab 3 also allows the analyst to perform several goodness-of-fit tests such as the chi-square test.

ExpertFit includes an option in tab 4 that allows one to display the representation of the inverted Weibull using different software packages. We show in Figure 4 the representations for four of the software packages supported by ExpertFit.

With some data samples, no candidate model provides an adequate representation. In this case we recommend the use of an empirical distribution function. One useful feature of ExpertFit is that in addition to using all of the sample values in the simulation software representation, it is possible to reduce the amount of required information through the use of a histogram-based empirical distribution function. We show a histogram-based representation (with 20 intervals) for two simulation software packages in Figure 5.

3 USING ExpertFit WHEN NO DATA ARE AVAILABLE

Quite often a simulation analyst must model a source of randomness for which no data are available. ExpertFit provides two analysis modes for this situation -- modeling of general activity times using triangular or beta distributions and modeling of random machine downtimes, for which we provide an example in this section. ExpertFit supports accurate modeling of systems with or without significant blocking or starving. For the example in this section, we will assume that the machine of interest is never blocked or starved.

Consider a machine that has an efficiency of 0.9; that is, it is actually producing parts 90 percent of the time. When the machine goes down, the average downtime is 60 minutes. However, the minimum downtime is 10 minutes. This information is specified to ExpertFit through a sequence of easy-to-use menus. After all of the required information has been specified, the average number of downs (actually the average number of busy-time/downtime cycles) per 8-hour shift is calculated by

Simulation Software	Representation
Arena	CONT(0.0000,24.800000, 0.0322,27.185000, 0.1576,29.570000, 0.3183,31.955000, 0.4791,34.340000, 0.5981,36.725000, 0.6945,39.110000, 0.7942,41.495000, 0.8457,43.880000, 0.8778,46.265000, 0.9068,48.650000, 0.9421,51.035000, 0.9550,53.420000, 0.9711,55.805000, 0.9807,58.190000, 0.9839,60.575000, 0.9904,62.960000, 0.9968,65.345000, 0.9968,67.730000, 0.9968,70.115000, 1.0000,72.500000)
AutoMod	continuous(0.0000:24.800000,0.0322:27.185000,0.1576:29.570000, 0.3183:31.955000,0.4791:34.340000,0.5981:36.725000,0.6945:39.110000, 0.7942:41.495000,0.8457:43.880000,0.8778:46.265000,0.9068:48.650000, 0.9421:51.035000,0.9550:53.420000,0.9711:55.805000,0.9807:58.190000, 0.9839:60.575000,0.9904:62.960000,0.9968:65.345000,0.9968:67.730000, 0.9968:70.115000,1.0000:72.500000)

Figure 5. Simulation Software Representation of the Empirical Distribution Function

Simulation Software	Busy-Time and Down-Time Representations
SIMSCRIPT II.5	GAMMA.F(540.000000, 0.700000, <stream>) 10.000000+GAMMA.F(50.000000, 1.400000, <stream>)
AweSim	GAMA(771.428571, 0.700000, <stream>) 10.000000 + GAMA(35.714286, 1.400000, <stream>)

Figure 6: Simulation Software Representations of Busy-Time and Downtime Models

the package to be 0.8. This makes sense since the average length of a busy-time/downtime cycle is 10 hours. A menu then allows various characteristics of the busy-time and downtime distributions to be displayed. We show the simulation software representations for two packages in Figure 6.

4 CONCLUSION

ExpertFit can help you develop more valid simulation models than if you use a standard statistical program, an input processor built into a simulation package (language or simulator), or hand calculations to determine input probability distributions. ExpertFit uses a sophisticated algorithm to determine the best-fitting distribution and, furthermore, has 43 built-in distributions. On the other hand, a typical simulation package contains roughly 10 distributions.

ExpertFit can represent most of its 43 distributions in 33 different simulation packages such as Arena, AutoMod, AweSim, COMNET III, FACTOR/AIM, GPSS/H, ManSim/X, MedModel, Micro Saint, MODSIM III, OPNET Modeler, ProModel, SES/*workbench*, SIMAN V, Simple++, SIMSCRIPT II.5, SLAMSYSTEM, Taylor II, and WITNESS, even though the distribution may not be available in the simulation package itself.

REFERENCES

Law, A.M. and W.D. Kelton. 1991. *Simulation Modeling and Analysis*, Second Edition. New York: McGraw-Hill.

Law, A.M. and M.G. McComas. 1989. Secrets of Successful Simulation Studies. *Industrial Engineering* 21: 28-31 (May 1989).

Law, A.M., M.G. McComas, and S.G. Vincent. 1994. The Crucial Role of Input Modeling in Successful Simulation Studies. *Industrial Engineering* 26: 55-59 (July 1994).

AUTHOR BIOGRAPHIES

AVERILL M. LAW is President of Averill M. Law & Associates, Inc. (Tucson, Arizona), a company specializing in simulation model building, training, and software. He has been a simulation consultant to more than 100 organizations, including General Motors, IBM, AT&T, General Electric, Nabisco, Xerox, NASA, the Air Force, the Army, and the Navy. He has presented more than 250 simulation short courses in 16 countries, and delivered more than 100 talks on simulation modeling at technical conferences.

He is the author (or coauthor) of three books and more than 35 papers on simulation, manufacturing, communications, operations research, and statistics, including the textbook *Simulation Modeling and Analysis* that is used by approximately 50,000 people worldwide. His series of papers on the simulation of manufacturing systems won the 1988 Institute of Industrial Engineers' best publication award. He is the codeveloper of the ExpertFit software package for selecting simulation input probability distributions, and he has developed several simulation videotapes. Dr. Law wrote a regular column on simulation for *Industrial Engineering* magazine from 1990 through 1991.

He has been a tenured faculty member and has taught simulation at the University of Wisconsin and the University of Arizona. Dr. Law has a Ph.D. in industrial engineering and operations research from the University of California at Berkeley.

MICHAEL G. MCCOMAS is Vice President for Consulting Services of Averill M. Law & Associates, Inc. He has considerable simulation modeling experience in a wide variety of application areas, and is the coauthor of seven published papers on simulation. His educational background includes an M.S. in systems and industrial engineering from the University of Arizona.