INTRODUCTION TO SIMULATION MODELING

J. William Schmidt

INTRODUCTION

Since World War II system modeling has played an increasingly important role in the analysis of complex systems in both the private and public sectors. In the broadest sense, a model may be considered to be a representation of reality without the presence of reality itself. Hence, pictures, graphs, management games, computer programs and mathematical equations may be considered models of those systems which they represent. For the purposes of this discussion we will restrict our attention to that class of models which attempts to capture the relationship between the behavior of a measure or measures of system effectiveness and the behavior of those variables and parameters which influence the measure(s) of effectiveness and includes simulation and mathematical models. The specific focus of our attention will be on simulation models.

SYSTEMS ANALYSIS AND MODELING

To analyze the behavior of a system under a variety of operating conditions the analyst may choose to experiment with the physical system itself or carry out his experiments on a model of the system. As experimentation with the physical system is usually disruptive to the functioning of the total organization or some part thereof, this alternative is usually infeasible or at least economically unattractive. At this point the analyst turns to analysis of the system through a model of the system.

The type of model chosen for the analysis is usually dependent upon the nature and complexity of the system and the capabilities of the analyst. In the analysis of organizational systems such as large corporations, governments, hospitals and the like, the system investigated is often represented through either a mathematical or a simulation model. The choice between a mathematical model and a simulation model depends upon the complexity of the system analyzed, the background and capability of the analyst and the history of success or failure of the two approaches to modeling within the organization.

The basic advantages of simulation modeling are versatility and simplicity. Many systems are sufficiently complex to defy a complete mathematical
description while being amenable to representation through a simulation model. In other cases, the system studied may be amenable to mathematical analysis but the level of mathematical sophistication required is beyond the background or capability of the analyst while he may possess the capability to develop a valid simulation model. That is, for reasonably complex systems the level of mathematical sophistication required for development of a valid mathematical model is usually more extensive than that required for development of the corresponding simulation model.

In general the steps taken in the analysis of a system are the same whether the analyst chooses to use a simulation model or a mathematical model. These steps can be summarized as follows:

1. Problem identification
2. Specification of the objectives of the analysis
3. Identification of the operating characteristics of the system and the collection of data describing the behavior of the system
4. Formulation of the system model
5. Parameter estimation
6. Preliminary model validation
7. Development of computer programs if necessary
8. Final model validation
9. Experimentation with the model
10. Analysis of results.

The primary focus of attention of this tutorial will be on steps 4, 5, 6, and 8.

HISTORICAL DEVELOPMENT

The origins of simulation may be traced to the sampling experiments of W. S. Gossett [18] (published under the name Student). However, the modern foundations of simulation are usually attributed to the work of von Neumann [21] and Ulam [20] in the
late 1940's which related to the analysis of nuclear shielding. The analytic technique employed was termed "Monte Carlo Analysis" and has since become fundamental to simulation modeling. Prior to the 1950's and the advent of the digital computer, most simulation experiments were carried out on analog computers. While analog computers are still used for this purpose, digital computers have become the dominant vehicle for execution of simulation experiments.

Although digital computers are discrete in nature, they are used to imitate the behavior of analog computers as well as in the analysis of discrete event systems. Today digital simulation is broadly accepted as a technique for the analysis of engineering, business, and behavioral systems and has been applied to problems such as:

The Analysis of Air Traffic Control Systems
The Analysis of Large-Scale Military Operations
The Analysis of the Traffic-Handling Capability of Telephone Switchboards
Job-Shop Scheduling
The Analysis of Forestry Operations
Analysis of the U.S. Economy
The Analysis of Multi-Product Inventory Systems
Determination of Manpower Requirements
Instructional Modeling for Higher Education
Competitive Market Analysis
Transportation Planning
Man-Machine Interface
Corporate Planning

FUNDAMENTALS OF DIGITAL SIMULATION
(Much of this material is taken from [16])

Simulation models may be classified as static or dynamic. A dynamic simulation model captures the time-dependent nature of the system under study. For example, the status of a waiting line system will change over time as customers arrive and are served and a dynamic simulation model of the system would reflect this time dependent variation. While most real world systems are in fact time dependent, situations exist where the time dependent nature of the system need not be captured for the purposes of analysis. Simulation models which do not take account of the time dependent nature of the system are referred to as static models. In the discussion which follows we will treat the development of both static and dynamic simulation models.

Static Simulation Modeling

Perhaps the simplest way to present the basic concepts of simulation modeling is by example. Consider a sampling system for quality control by attributes. Manufacturing lots containing L items are submitted for inspection. The inspection procedure consists of drawing a sample of size n, inspecting each item in the sample, identifying each as either good or bad, recording the total number of defects found, x, and comparing the number of defects identified with a criterion variable called the acceptance number, c. If the number of defects found in the sample is less than or equal to the acceptance number the lot is accepted. Otherwise, the lot is rejected. Let us assume that the objective of the analysis is to determine the proportion of lots which one might expect to be rejected as a result of implementation of the quality control system.

First let us examine how one might determine the proportion of lots rejected by experimenting with the physical system. This could be accomplished by implementing the inspection system defined above and using it for the inspection of M manufacturing lots. Associated with each lot selected is the number of items contained in the lot and the proportion of those items which are defective. For simplicity we will assume that the lot size is constant from one lot to another. However a similar assumption with respect to the proportion of defective items contained in each lot would be unrealistic. Thus we must assume that an unknown proportion of defective items will be contained in each manufacturing lot, and that the proportion defective will vary from lot to lot. From each lot we draw a sample of size n and carry out the sampling procedure already defined. Repeating this process for a total of M lots we record the total number of lots rejected and divide this number by the total number of lots inspected, N, to obtain an estimate of the proportion of lots rejected. A schematic representation of the experiment with the physical system is shown in Figure 1.a.

Now let us consider the development of a simulation model to conduct a similar analysis. It is relatively simple to develop a computer program to execute the steps indicated in Figure 1.a with the exception of definition of the proportion defective for each lot and the execution of inspection of each item included in the sample. Thus to completely define the simulation model we must develop a method for assigning a value to the proportion of defective items included in the lot such that the simulated variation in proportion defective from one lot to another is representative of the variability which exists in the real world system. In addition, complete specification of the simulation model will require the development of a technique whereby each item in the sample is classified as good or bad in a manner descriptive of actual conditions.

Let us treat the problem of simulating the inspection of individual items in the sample first. It should be obvious that the proportion of defective items in the lot will influence the number of
FIGURE 1
Flowcharts for Physical (a) and Simulated (b) Sampling Inspection Systems
defective items detected in the sample. We will assume at this point that a value has been assigned to the proportion of defective items in the lot, P. If an item is drawn at random from a lot having a proportion of defective items P, then the probability that the item is defective is P. Hence the methodology developed for simulation of the inspection process should have the property that the probability that any item selected is defective is P as it is in the case of the real world system. To accomplish this we will use what are known as "random numbers". A random number is a random variable which is uniformly distributed on the interval (0,1). Hence, in drawing or generating a random number each number between 0 and 1 has an equal and independent chance of occurring.

Let us examine how one would use a random number to determine whether or not an individual item of product is defective. Suppose that the proportion of defective items in a lot is 0.05, that is P = 0.05. If we were to draw a succession of items from the lot we would expect to find that approximately 5% were defective. Now let us draw a sequence of random numbers. Since these numbers have the property that each value between 0 and 1 has an equal and independent chance of occurrence, we would expect approximately 5% of the numbers drawn to lie on the interval (0.0, 0.05).

Thus to simulate the inspection of items, we draw a sequence of n random numbers and compare each to the proportion defective, P. If the random number, which we shall designate by r, is less than or equal to P, we will classify the item as defective. On the other hand, if r is greater than P we will designate the item as good. To put the process in probabilistic terms, we will designate a 0 as corresponding to a bad item and 1 as corresponding to a good item. The probability of a 0 occurring (bad item) is then P and the probability of a 1 (good item) is 1−P. The cumulative probability distribution for this random variable is shown in Figure 2. By choosing a random number, r, we are actually designating a value of the distribution function. Entering the y axis in Figure 2 at the point designated by the random number we proceed horizontally until we intersect the distribution function. Dropping vertically from the point of intersection we arrive at the value assigned to the random variable. As indicated previously, if r is less than or equal to P, a 0 is generated and if r is greater than P, a 1 is generated. Hence, we have devised a synthetic method of categorizing items as good or bad which has the property that the probability that any item is categorized as defective is the same as the probability that the item would be found defective in the physical inspection process.

The fundamental methodology used to generate all random variables is similar to that described for the inspection process discussed above. Let us apply this approach to the generation of values of proportion defective for those lots for which the inspection process is to be simulated. Proportion defective is a random variable which can assume values between 0 and 1. A typical cumulative distribution function for proportion defective is shown graphically in Figure 3. To generate values of proportion defective we select a random number, r, enter the y axis of the distribution function at the point designated by r, proceed horizontally until we intersect the distribution function, and drop down at this point to the x axis to pick up the value of proportion defective generated.
The reader will recall that the two central problems which existed in developing a simulator for the quality control system discussed here were to identify synthetic means of generating proportion defective in a manner such that the variability in the values of proportion defective generated would correspond to those which exist in the real world system and to devise a method of designating sampled items as good or bad. Having developed these two techniques we are in a position to complete the logic for preparation of the final simulation model. A flowchart for the simulation model for this quality control system is shown in Figure.

In discussing the development of the simulator for the sampling inspection system described above, we have assumed that the analyst has available a source of random numbers and that he is able to construct the cumulative probability distribution functions required for generation of the random variables included in the system. Random numbers can be obtained from standard tables or can be generated on a digital computer. We will discuss random numbers further in a later section of this paper. In order to construct the cumulative distribution function of a random variable data must be collected from the system under study which indicates the variation which exists in the random variable of interest. While the data collection effort required may be expensive and time consuming, it is necessary prerequisite to the development of a valid simulation model or any system model for that matter.

**Dynamic Simulation Modeling**

Dynamic or time dependent simulation models take account of status change in systems as they occur through time. Keeping track of events and changes in system status over time is generally accomplished in one of two ways. The first and perhaps most widely used method for tracking system behavior over time is through the next event approach. In the next event method the simulation model keeps track of the times at which each distinct status changing event will occur. It then chooses the time at which the next event will occur, moves forward in the simulation to that point in time, and alters the status of the system in accordance with the conditions dictated by the occurrence of the event. The philosophy underlying the next event method is that there is no need to view the system at points in time other than those at which status changing events occur. An alternative to the next event method is the fixed time step approach. Using the fixed time step method the simulation model moves forward in time in constant increments recording those status changing events which have occurred since the last step was taken and altering the status of the system accordingly. While the next event method is generally more precise than the fixed time step method, the precision of the fixed time step method can be increased by reducing the time increment for each step.

Simulation of time dependent systems will be introduced by example. For this example the next event method will be used for time keeping. Consider a system of 5 production lines which are to be maintained by a single repair crew. From time to time the production lines fail and repairs must be carried out. However at the time a given line fails, the repair crew may be busy with another production line which has failed previously. In this case the failing production line would have to wait for service until the repair crew becomes available. It will be assumed that the repair crew services production lines in the order in which they fail, that is first-come, first-served, and that the production system operates three shifts per day five days per week. For the purposes of this example we will assume that the objective of the analysis is to develop a simulation model which can be used to estimate the failure rate, average down time, average repair time, and average time waiting for service per week for each of the five production lines.

In developing a simulation model for this system one must recognize that there are ten events which may change the status of the system. First a failure on any one of the five production lines will alter the status of the system as will the completion of a repair on any line. In addition the end of the period of simulation must be recognized by the model in order to terminate the analysis. The simulation analysis moves forward in time by comparing the times of occurrence of each of the 11 events just mentioned. The model selects the time at which the next event will occur and moves forward in time to that point, altering the status of the system as dictated by the occurrence of the next event. In order to compare the times of occurrence of each of these 11 events the simulation model must constantly have available a recorded time for each.

Let us assume that all five production lines are operating when the period of simulation begins. Using methods similar to those already described, the simulation model generates and records the time until the first failure for each of the production lines. Until the simulation model moves forward in time to the point of the first line failure, a time of service completion is not generated for any line. However the reader will recall that the simulation model determines the next event in the simulation by comparing the times recorded for failure of each of the five production lines, repair of each of the five production lines, and the point in time at which the simulation is to end. However, at the beginning of the simulation experiment no times have been generated or recorded for completion of the repair to any production line, since production lines have not yet failed. Hence, the next event in the simulation must be either the failure of a production line or the end of the period of simulation. To make sure that the simulation model picks up one of these six events as the next event, at the beginning of the simulation experiment we record the time of the next repair for each line as infinity, or more appropriately as some arbitrarily large value which exceeds the duration of the simulation experiment. As we move forward in the simulation, a point in time will be reached at which a production line fails. At this point the simulation model will record the production line as down. Since the production line cannot fail again until it is serviced, the time of the failure for this production line will be recorded as infinity to avoid the problem of picking up a second failure on a line which is already down. In general, then, we assign an arbitrarily large time to any event which
Introduction to Simulation Modeling (Continued)

logically cannot be the next event in the simulation.

Let us now consider the operation of the simulation model at an arbitrary point in the simulation experiment. The simulation model searches through the times of occurrence of each of the 11 possible events which might take place next. Let us assume that the next event in the simulation is the failure of a production line. The simulation model moves forward in time to the point in which this failure becomes effective, records the failure time, increases the number of lines down by one, accumulates the number of failures which have taken place on this line thus far, and sets the time to the next failure to infinity. Since the repair crew may not be available for servicing this line, the simulation model must determine whether or not the repair crew is busy servicing another production line. If the repair crew is unavailable for service of the line which has just failed, that production line is placed in the waiting line. Otherwise, the repair crew is assigned to the failing production line, a repair time for the line is generated and the point in time at which the repair will be completed is computed.

Now let us consider the case where the next event in the simulation is the completion of service on one of the production lines. In this case the number of lines down is reduced by one, the repaired production line is put back in service, the time until the next failure for that line is generated, the point in time at which the failure of this line will occur is calculated, the time of the next repair on this line is set to infinity, and repair time and down time for the line are accumulated. Since the repair crew has just completed service on one of the production lines, the simulator must determine whether another production line is waiting for service. If a production line is waiting for service then the repair crew is assigned to that production line, service time is generated, and the cumulative waiting time and down time for that line are increased. If there are no further production lines waiting for service, the repair crew is placed in an idle status.

When the next event in the simulation experiment is the end of the period of simulation, summary statistics must be calculated and printed out. In this case the simulation model would compute the failure rate, average down time, average waiting time, and average repair time per week for each of the production lines and print this information out for the analyst. The simulation model for this system is shown in the flowchart in Figure 4. A summary of the results of the simulation of this system for a period of 20 weeks is given in Table 1. Similar results are shown in Table 2 for a 20 week period of observation of the system. The reader will note that the results of the 20 week period of simulation are quite similar to those obtained for the 20 week period of observation. It is a check of this type which can serve as a means to validating the simulation model developed. However, in general the analyst should compare simulated results with those obtained from observation of the physical system through appropriate statistical testing procedures rather than by simple visual comparison.

RANDOM NUMBERS

As the two examples cited above indicate, random number generation is an important component of every stochastic simulation model. An essential property of every random number generator is the ability to generate random variables which are uniformly distributed on the interval (0,1). Actually digital methods for generating random numbers are algorithmic and therefore the numbers resulting are generally termed pseudo-random numbers. That is, since the numbers are generated algorithmically they are not actually random. However, were one to compare a set of numbers derived from a reliable digital generator with numbers which were truly random, the distinction between the two sets of numbers would not be apparent. A discussion of algorithmic methods for the generation of random numbers is beyond the scope of this article. However, random number generators are available for most high-speed digital computers.

GENERATION OF RANDOM VARIABLES

At the heart of every simulation model is a mechanism for generating values of those random variables which influence the behavior of the system analyzed. The method used to generate values of a random variable is often referred to as a process generator. Fundamentally, a process generator defines a relationship between each possible value of the random variable considered and values of a uniformly distributed random number. The principle underlying the generation of random variables was discussed in the preceeding sections and is illustrated graphically in Figures 2 and 3.

While the graphical approach to process generation is acceptable in some cases, it is often useful to define mathematical relationships which simplify the process. The reader will recall that in the graphical approach we selected the value of a uniformly distributed random number which was in turn used to define a specific value of the cumulative distribution function of the random variable to be generated. Let \( r \) be the value of the random number and let \( F(x) \) be the value of the distribution function such that

\[
F(x) = r
\]

where \( r \) and \( F(x) \) lie on the interval \((0,1)\). The problem at hand is to find the value of the random variable \( x \) which satisfied (1). That is, \( x \) is related to \( r \) through (1) and we must find an inverse relationship of the form

\[
x = h(r)
\]

To illustrate how this is accomplished, assume that \( x \) is exponentially distributed with the probability density function given by

\[
f(x) = \lambda e^{-\lambda x}, \quad 0 < x < \infty
\]
Flowchart for the Maintenance Simulator

START

Read input data
Duration of simulation period
Parameters of failure and service distributions

Initialize counters and accumulators
For each production line:
Total no. of failures = 0
Total down time = 0
Total repair time = 0
Total wait time = 0
No. lines currently down = 0

Generate time until failure for each production line
Set time of next service completion for each line to $\infty$

Is next event a line failure?

Yes

Determine which line fails
Record failure time
Increase no. of lines down by one
Set time of next failure on this line to $\infty$
Accumulate failures on this line

No

Reduce no. of lines down by one
Determine which line has completed repair
Generate time until next failure for repaired line
Calculate time of next failure by repaired line
Set time of next repair on this line to $\infty$
Accumulate repair time and down time for repaired line

Is repair crew available?

Yes

Assign repair crew to failing line
Record time service begins
Generate repair time
Calculate time of repair completion

No

Place failing line in waiting line for service

End of simulation
Accumulate down time, repair time, and waiting time for each line

Write output for each line
Average no. failures, average down time, average waiting time, average repair time

Stop

Is another line waiting for repair?

Yes

Find first production line in wait line
Accumulate waiting time and down time for that line

No

Repair crew has completed service and is idle

FIGURE 4
The distribution function of \( x \), \( F(x) \), is given by
\[
F(x) = 1 - e^{-\lambda x} \quad (4)
\]
Now when we choose or generate a value of \( r \) we are simply specifying a particular value of \( F(x) \). That is
\[
r = F(x) = 1 - e^{-\lambda x} \quad (5)
\]
Solving for \( x \) in terms of \( r \) we have
\[
x = -\frac{1}{\lambda} \ln(1-r) \quad (6)
\]
Thus we have a simple mathematical expression which will define a unique value of the exponential random variable \( x \) for each value of \( r \) generated.

Although conceptually the technique described above may be used for the development of a process generator for any random variable, its application may lead to a tedious exercise in the case of many random variables. For example, consider the normal random variable with probability density function given by
\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-u)^2}{2\sigma^2}}, -\infty < x < \infty \quad (7)
\]
The procedure outlined above calls first for development of the distribution function of \( x \), given by
\[
F(x) = \int_{-\infty}^{x} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-u)^2}{2\sigma^2}} dx \quad (8)
\]
However, in this case the integral in (8) must be evaluated numerically which in turn leads to substantial difficulties in deriving a solution for \( x \) in terms of a random number \( r \).

Even in the case where a simple expression for the distribution function can be achieved, the inversion technique discussed so far may prove cumbersome. To illustrate suppose that \( x \) is Erlang distributed with probability density function
\[
f(x) = 2\lambda e^{-2\lambda x} x, \quad 0 < x < \infty \quad (9)
\]
The distribution function of \( x \) is given by
\[
F(x) = \frac{1}{2\lambda} \left[ 1 - e^{-2\lambda x} - 2 \lambda xe^{-2\lambda x} \right] \quad (10)
\]
At this point we would normally set the left-hand-side of (10) equal to a random number \( r \) and solve for \( x \) in terms of \( r \). However, in this case the solution for \( x \) can be achieved only through numerical methods.

Cases may be encountered where the value of a random variable \( x \) associated with a random number \( r \) can be identified only through numerical methods. However it is often useful to attempt to identify a relationship between the random variable of interest and other random variables which can be generated conveniently. For example, the chi square random variable with \( n \) degrees of freedom can be expressed as the sum of the squares of \( n \) standard normal random variables. The Erlang random variable may be expressed as the sum of identically distributed exponential random variables. In a similar fashion the binomial random variable may be expressed as the sum of Bernoulli random variables and the negative binomial random variable may be expressed as the sum of geometric random variables.

**SYSTEMS ANALYSIS THROUGH SIMULATION**

Thus far our discussion of simulation methodology has dealt primarily with the mechanics of the development of simulation models. We have said very little about how such models can be used as an analytic tool and as an aid to the decision making process. An in depth discussion of these topics is far beyond the scope of this paper. However, there are certain critical aspects with respect to systems
analysis through simulation that the reader should be aware of. We will simply introduce the reader to the two most important of these problem areas without attempting to deal with their solution.

In attempting to develop a simulation model one of the first difficulties to be dealt with is that of model validation. Before a simulation model can be used for systems analysis the analyst must first verify that the model is an adequate reflection of reality. This can sometimes be achieved by comparing model results with those of real world system. However, even though a model realistically reflects the performance of a system in the past and today, there is no guarantee it will do so under conditions which may prevail in the future. All too often the analyst has no objective basis for ascertaining model validity but rather must rely upon his own judgement and that of others familiar with the system under study.

Simulation models are often used to estimate a measure of measures of performance of the system studied. However, unlike mathematical models, the output of a simulation model is one or more random variables. Hence, there is error in an estimate of system performance through simulation. This is both an advantage and a disadvantage. Since the output of the real world system is also a random variable, the random nature of the output of the simulation model might be considered a realistic reflection of true system performance making no attempt to hide the difficulties which exist in assessing the value of that measure. On the other hand, suppose that we wished to compare several systems based upon the mean value of an appropriate measure of system performance. In this case we require an estimate of system performance for each system which is precise enough to validly carry out the comparison indicated. The precision with which such measures of performance are estimated can usually be improved by increasing the length of the simulation run. However, as the simulation run length is increased the cost of the analysis also increases. Thus, the analyst is faced with the problem of balancing the cost of the analysis against the precision of estimation to be achieved.

**SUMMARY**

Simulation has proven to be an effective and versatile modeling technique for the analysis of complex interactive systems in both the private and public sectors. In addition simulation offers the advantage of relative simplicity in model construction as compared with mathematical modeling although the analysis of simulated results may require an advanced knowledge of probability theory and statistical methodology.

Although the advantages of simulation are substantial, it should not be used indiscriminately. In general, if a system can be modeled mathematically it should be. A mathematical model can usually be evaluated more quickly than a simulation model since replication of a given system condition is normally required for reliable estimation of the measure of system performance when the system is modeled through simulation. Thus the cost of the analysis is likely to be greater when the system is modeled through simulation than through a mathematical model.

**REFERENCES AND SUGGESTED READING**


