ELSEVIER

European Journal of Operational Research 82 (1995) 145–162

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

Theory and Methodology

Verification and validation of simulation models

Jack P.C. Kleijnen

CentER and Department of Information Systems and Auditing, Katholieke Universiteit Brabant (Tilburg University), P.O. Box 90153, 5000 LE Tilburg, Netherlands

Received September 1992; revised November 1993

Abstract

This paper surveys verification and validation of models, especially simulation models in operations research. For verification it discusses 1) general good programming practice (such as modular programming), 2) checking intermediate simulation outputs through tracing and statistical testing per module, 3) statistical testing of final simulation outputs against analytical results, and 4) animation. For validation it discusses 1) obtaining real-world data, 2) comparing simulated and real data through simple tests such as graphical, Schruben–Turing, and t tests, 3) testing whether simulated and real responses are positively correlated and moreover have the same mean, using two new statistical procedures based on regression analysis, 4) sensitivity analysis based on design of experiments and regression analysis, and risk or uncertainty analysis based on Monte Carlo sampling, and 5) white versus black box simulation models. Both verification and validation require good documentation, and are crucial parts of assessment, credibility, and accreditation. A bibliography with 61 references is included.

Keywords: Simulation; Statistics; Regression; Risk analysis; modelling

1. Introduction

Terminology in the area of verification and validation or V&V is not standard; see Barlas and Carpenter (1990, p.164, footnote 2), Davis (1992a, p.4), and Murray-Smith (1992). This paper uses the definitions of V & V given in the classic simulation textbook by Law and Kelton (1991, p.299): "*Verification* is determining that a simulation computer program performs as intended, i.e., debugging the computer program.... *Validation* is concerned with determining whether the conceptual simulation model (as opposed to the computer program) is an accurate representation of the system under study". Therefore this paper assumes that verification aims at a 'perfect' computer program, in the sense that the computer code has no programming errors left (it may be made more efficient and more user friendly). Validation, however, can not be assumed to result in a perfect model, since the perfect model would be the real system itself (by definition, any model is a simplification of reality). The model should be 'good enough', which depends on the goal of the model. For example, some applications need only relative (not absolute) simulation responses corresponding to different scenarios; see Section 3.3.

Another well-known author on V & V in simulation discusses these issues for the various *phases* of modeling: Sargent (1991, p.38) states "the *conceptual model* is the mathematical/logical/verbal representation (mimic) of the problem entity developed for a particular study; and the *computer-*

SSDI 0377-2217(94)00016-6

ized model is the conceptual model implemented on a computer. The conceptual model is developed through an *analysis and modelling phase*, the computerized model is developed through a *computer programming and implementation phase*, and inferences about the problem entity are obtained by conducting computer experiments on the computerized model in the *experimentation phase*". The conceptual model is also discussed in detail by Oral and Kettani (1993).

In practice V&V are important issues. A computer program with bugs may generate output that is sheer nonsense, or worse, it may generate subtle nonsense that goes unnoticed. A nonvalidated model may lead to wrong decisions. In practice, verification and validation are often mixed; see Davis (1992a, pp.5–6) and also Miser (1993, p.212).

The interest in V&V shows a sharp increase in the USA defense community; see Davis (1992 a,b), Fossett, Harrison, Weintrob, and Gass (1993), Pace (1993), Pacheco (1988), Williams and Sikora (1991), and Youngblood (1993). In Europe and China the defense organizations also seem to take the initiative; see Kleijnen and Alink (1992) and Wang, Yin, Tang and Xu (1993). The renewed interest in V&V is also illustrated by the publication of a monograph on validation by Knepell and Arangno (1993) and the Special Issue on "Model Validation in Operational Research" of the *European Journal of Operational Research*; see Landry and Oral (1993).

There is no standard theory on V&V. Neither is there a standard 'box of tools' from which tools are taken in a natural order; see Davis (1992a, p.19) and Landry and Oral (1993). There does exist a plethora of philosophical theories, statistical techniques, software practices, and so on. Several *classifications* of V&V methods are possible; examples are provided by Davis (1992a), Fossett et al. (1991), Landry and Oral (1993), Oral and Kettani (1993), Pace (1993), and Williams and Sikora (1991). The emphasis of this article is on *statistical techniques*, which may yield reproducible, objective, quantitative data about the quality of simulation models. To classify these techniques, the paper stresses that in practice the quantities of data on simulation inputs and outputs may vary greatly; also see Bankes (1993), Oral and Kettani (1993, p.223) and Wang et al. (1993). The objective of this paper is to survey statistical V&V techniques. Moreover, it introduces two new statistical techniques for validation (based on familiar regression analysis).

Unfortunately, it will turn out that there are no perfect solutions for the problems of V&V in simulation. The whole process has elements of art as well as science (the title of one of the first books on simulation was *The Art of Simulation*; see Tocher, 1963). Taking a wider perspective than simulation, Miser (1993, p.207) states: "The nature of scientific inquiry implies that it is impossible to eliminate pitfalls entirely"; also see Majone and Quade (1980).

These problems occur in all types of models (for instance, econometric models) and in all types of computer programs (for example, bookkeeping programs), but this paper concentrates on simulation models in operations research. (Expert systems or more generally, knowledge based systems are closely related to simulation models; their validation is discussed in Benbasat and Dhaliwal (1989); also see Davis (1992a).)

This article is organized as follows. Section 2 discusses verification. Section 3 examines validation. Section 4 briefly reviews documentation, assessment, credibility, and accreditation. Section 5 gives supplementary literature. Section 6 provides conclusions. It is followed by a list of 61 references. (To avoid dragging along a cumulative list of everything published on V&V in simulation, only those publications are included that either seem to deserve special mention or that are not mentioned in the references of this paper. This paper includes three bibliographies, namely Balci and Sargent (1984a), DeMillo, McCracken, Martin and Passafiume (1987), and Youngblood (1993).)

2. Verification

Once the simulation model has been programmed, the analysts/programmers must check if this computer code contains any programming errors ('bugs'). Several techniques are applicable,

but none is perfect. This paper discusses 1) general good programming practice such as modular programming, 2) checking of intermediate simulation outputs through tracing and statistical testing per module, 3) comparing (through statistical tests) final simulation outputs with analytical results, and 4) animation.

2.1. General good programming practice

Software engineers have developed numerous procedures for writing good computer programs and for verifying the resulting software, in general (not specifically in simulation). Software engineering is indeed a vast area of research. A few key terms are: modular programming, object oriented programming, chief programmer's approach, structured walk-throughs, correctness proofs. Details are given in Adrion, Branstad and Cherniavsky (1982), Baber (1987), Dahl (1992), DeMillo et al. (1987), and Whitner and Balci (1989); also see Benbasat and Dhaliwal (1989) and Davis (1992a). A comprehensive bibliography can be found in DeMillo et al. (1987).

Modular testing will be further discussed in the next subsections. Object orientation was already implemented in the old simulation language Simula 67. The importance of good documentation for both verification and validation will be discussed in Section 4.

2.2. Verification of intermediate simulation output

The analysts may calculate some intermediate simulation results *manually*, and compare these results with outputs of the simulation program. Getting all intermediate results from a computer program automatically is called *tracing*. Even if the analysts do not wish to calculate intermediate results by hand, they can still 'eyeball' the program's trace and look for programming errors. Davis (1992a, pp.21–23) seems to equate 'eye balling' with 'face validity'. Modern simulation software provides tracing facilities and more advanced 'debuggers'; see Pegden, Shannon and Sadowski (1990, pp.137–148).

In practice, many simulation programs are very big. Good programming requires that the computer code be designed *modularly* (no 'spaghetti programming'; see Section 2.1 and Davis, 1992a, p.23). Then the analysts 'divide and conquer', that is, they verify the total computer code, module by module. Different members of the team may check different modules. Some examples now follow.

1) The analysts may test the *pseudorandom number generator* separately, if they had to program that generator themselves or they do not trust the software supplier's expertise. By definition, random numbers are continuous statistical variables, uniformly distributed between zero and one, and statistically independent. The main problem in practice is that pseudorandom number generators give outputs that are not independent (but show a 'lattice structure'). Selecting a new generator may result in better statistical behavior. Moreover the pseudorandom number generator may be wrong because of programming errors: many generators require either machine programming or rather sophisticated programming in a higher language.

Schriber (1991, p.317) points out that GPSS/H automatically computes chi-square statistics to test the hypothesis that the pseudorandom numbers used in a particular simulation experiment, are uniformly distributed. Ripley (1988, p.58) mentions two simulation studies that gave wrong results because of an inferior generator. Kleijnen and Van Groenendaal (1992) provide a detailed discussion of different types of pseudorandom number generators and of many tests to verify their correctness.

2) The analysts may further test the subroutines that generate samples from certain *non-uniform distributions*. Experience shows that analysts may think that the computer gives normal variates with standard deviation (say) 10, whereas actually the variates have a *variance* of 10. This confusion is caused by the lack of standard notation: some authors and some software use the notation $N(\mu, \sigma)$, whereas others use $N(\mu, \sigma^2)$. Similar confusion arises for exponential distributions: some authors use the parameter (say) λ to denote the mean interarrival time, but others use that symbol to denote the arrival rate.

The analysts may also specify the wrong unit of

measurement, for instance, seconds instead of minutes. In this example the results are wrong by a factor 60.

To verify that the random variate subroutine does what it is intended to do, the analysts should first of all read the documentation of the subroutine. Next they may estimate the mean and variance of the sampled variable, and compare those statistics with the theoretical values. These values are indeed known in a simulation study; for instance, service times are sampled from an exponential distribution with a known mean, namely the mean that is input to the simulation program. Systematic deviations between the observed statistics and the theoretical values may be detected through parametric or through distribution-free tests. An example of a t test will be discussed in Eq. (4).

Random (not significant, not systematic) deviations between the sample average (say) $\bar{y}$ and its expected value μ_y always occur (random variables are underlined). To reduce the effect of such a deviation, a variance reduction technique (VRT) called *control variates* can be applied. This VRT corrects $\underline{x}$, the simulation output (for example, average waiting time), for the random deviation between the input's sample average and population mean:

$$\underline{x}_c = \underline{x} + \beta(\mu_y - \underline{\bar{y}}), \tag{1}$$

where a proper choice of the coefficient β means that the variance of the new estimator $\underline{x}_c$ is reduced. See Kleijnen and Van Groenendaal (1992, pp.200–201).

Instead of testing only the mean or variance, the analysts may test the whole distribution of the random variable. Then they can apply a goodness-of-fit test such as the well-known chi-square and Kolmogorov–Smirnov tests; see the survey in Kleijnen (1987, pp.94–95).

2.3. Comparing final simulation outputs with analytical results

2.3.1. Introduction

The *final* output of (say) a queueing simulation program may result only after millions of customers have been processed. This is indeed the case if the steady state mean waiting time is of interest and traffic intensity is high. Another example is provided by the simulation of 'rare events' such as breakdowns of highly reliable systems. Verifying such types of simulation responses by hand or by eyeballing the trace (discussed in the preceding subsection) is practically impossible. Restricting attention to *short* time series is misleading.

In these situations the analysts may verify the simulation response by running a *simplified* version of the simulation program with a *known analytical solution*. This approach assumes that the analysts can indeed find a 'test case' with a known solution, but this is not an unrealistic assumption. For example, in logistics simulation the analysts often model reality as a queueing system. Then the analysts can use a textbook on queueing theory to find formulas for the steady state expectations of several types of response (mean waiting time of jobs and mean utilizations of machines). These formulas, however, assume Markovian (exponential) arrival and service times, with (say) n servers: M/M/n models. First the analysts can run the simulation program with exponential arrival and service times, only to verify the correctness of the computer program. Suppose the response of that simulation does not significantly deviate from the known mean response (see the statistical test in Eqs. (2)–(4) in Section 2.3.2). Next they run the simulation program with non-exponential input variables to simulate the responses that are of real interest to the users. The analysts must then hope that this minor change in the computer program does not introduce new bugs.

It may be asserted that in all simulation studies the analysts should be guided by knowledge of theoretical models with known solutions, when they study real systems. In many simulation studies the analysts model reality as a (complicated) queueing system. There is much literature on queueing systems. These systems comprise servers, in parallel and in sequence, and customers who can follow different paths through the queueing network. For certain queueing networks (for example, with infinite buffers for work in process) steady state solutions can be com-

puted numerically. Besides numerous textbooks and articles there is software that gives analytical, numerical, and simulation results; see Kleijnen and Van Groenendaal (1992, p.127). Indeed much research is going on in queueing theory with applications in computer, communications, and manufacturing systems. In other areas (for example, inventory management and econometrics) there is also a substantial body of theory available; see Kleijnen and Van Groenendaal (1992). In a mine hunting case study there is an analytical model besides a simulation model; see Kleijnen and Alink (1992). The importance of 'theoretical analysis' is also discussed in Davis (1992a, pp.18–19). So a stream of publications and software can help the simulation analysts to find models that are related to their simulation models and that have analytical or numerical solutions. General systems theory emphasizes that the scope of a study can be reduced by either studying a subsystem only (say, queueing at one specific machine) or by restricting the response types (for example, financial variables only); also see Davis (1992b). In this way the analysts may find simplified models with known responses for certain modules or they may verify certain response types of the total simulation program.

Simulating a related system with known solution may also be used to reduce the variance through control variates. Now in (1) $\bar{y}$ denotes the average response of the simulated system with known response, μ_y denotes the known expected value of that response, x is the simulation response of real interest, x_c is the better estimator, both systems are simulated with common pseudorandom numbers. The more the two systems are similar, the higher is the correlation between their responses and the lower is the variance of the new estimator for the system of real interest. Also see Kleijnen (1974, pp.162–163).

So the effort of simulating a related system with known solution *may* pay off, not only in debugging but also in variance reduction through control-variates. But there are no guarantees!

In some situations no mathematical statistics is needed to verify the correctness of the simplified simulation model, namely if that model has only *deterministic* inputs (so the simplified simulation is deterministic whereas the simulation model of real interest may be random). One example is an inventory model with constant demand per period, so – under certain other assumptions – the classic 'economic order quantity' (EOQ) solution holds. A second example is a single server queueing model with constant arrival and service times (say) $1/\lambda$ and $1/\mu$ respectively with $\lambda/\mu < 1$, so it is known that the utilization rate of the server is λ/μ and that all customer waiting times are zero. Examples of economic models with deterministic inputs and known outputs are given in Kleijnen and Van Groenendaal (1992, pp.58–64). In these examples the simulation responses must be identical to the theoretical responses (except for numerical inaccuracies).

2.3.2. Statistical technique

How can analysts compare the output of the simplified simulation program with its known expected value? They should understand that in the steady state the system is still stochastic (but the probability law that governs the stochastic process no longer depends on the initial state), so mathematical statistics is needed. Hence they should use a statistical test to verify that the expected value of y, the simulation response of the *simplified* simulation program, is equal to the *known* steady state mean μ_y:

$$H_0: E(y) = \mu_y. \tag{2}$$

The well-known Student t test assumes normally and independently distributed (NID) simulation responses y with mean μ_y and variance σ_y^2. To estimate this unknown variance, the analysts may partition the simulation run into (say) m subruns and compute y_i, the average of subrun i, and $\bar{y}$, the average of these m subrun averages (which is identical to the average of the whole simulation run), which yields

$$s_y^2 = \sum_{i=1}^{m} \frac{(y_i - \bar{y})^2}{m-1}. \tag{3}$$

Then the test statistic becomes

$$t_{m-1} = \frac{\bar{y} - \mu_y}{s_y/\sqrt{m}}. \tag{4}$$

Many simulation responses are indeed approximately normally distributed: a variation of the central limit theorem applies, when the simulation response is the average of autocorrelated waiting times of successive customers. If the simulation response is not (approximately) normal, then the t test may still be applied because this test is not very sensitive to nonnormality, especially if m is large; see Kleijnen (1987, pp.14–23).

(Kleijnen and Van Groenendaal (1992, pp. 190–195) present several alternative approaches (such as renewal analysis) to the estimation of the variance of the simulation response in the steady state. Kleijnen (1987, pp.23–25) discusses several distribution-free tests.)

In practice, however, most simulation studies concern the behavior of the real system in the *transient* state, not the steady state. For example, the users may be interested in the total waiting time during the next day – under various scheduling algorithms (priority rules) – so the simulation run stops as soon as the end of that simulated day is reached. Such types of simulation are called 'terminating' simulations. When verifying such a simulation, there are usually no analytical or numerical solutions available: most solutions hold in the steady state only. The analysts may then first simulate a non-terminating variant of the simulation model, for verification purposes only. Next they change the simulation program, that is, they introduce the terminating event (in the example this event is the 'arrival' of the end of the working day). As pointed out (in Section 2.3.1, paragraph 2), they must then hope that this minor change in the computer program does not introduce new bugs. Again, there is no guarantee (see Section 1).

There is a *statistical* complication, as virtually all simulation programs have *multiple* responses (for example, mean waiting time of jobs and mean utilizations of machines). So the computer program transforms (say) S inputs into T outputs with $S \geq 1$ and $T \geq 1$. That transformation must be correct for all response types of the simplified simulation program with known means. Consequently the probability of rejecting a null-hypothesis like (2) increases as T (the number of responses) increases, even if the program is correct. This property follows from the definition of the type I or α error of a statistical test (different error types will be further discussed in Section 3.2). Fortunately there is a simple solution based on *Bonferroni's inequality*. Traditionally the t_{m-1} value in (4) is compared with $t_{m-1;\,\alpha/2}$, which denotes the critical value taken from the table for the t statistic with $m-1$ degrees of freedom, type I error probability fixed at α, in a two-sided test. Using Bonferroni's inequality, the analysts merely replace α by α/T. This implies that bigger discrepancies between the known means and the simulation responses are accepted:

$$t_{m-1;\,\alpha/2} < t_{m-1;\,\alpha/(2T)}.$$

It can be proved that Bonferroni's inequality keeps the overall 'experimentwise' error probability below the value α. It is recommended to combine the Bonferroni inequality with a value such as $\alpha = 0.20$ instead of the traditional value 0.05.

(Multivariate techniques provide alternatives to this combination of univariate techniques (such as the t test in Eq. (4)) and Bonferroni's inequality. Multivariate techniques are more sophisticated, but not always more powerful; see Balci and Sargent (1984b), Barlas (1990), and Kleijnen and Van Groenendaal (1992, pp.144,155).)

2.4. Animation

To verify the computer program of a dynamic system, the analysts may use *animation*. The users then see dynamic displays (moving pictures, cartoons) of the simulated system. Since the users are familiar with the corresponding real system, they can detect programming errors (and conceptual errors too, but that concerns validation). Well-known examples are simulations that show how vehicles defy the laws of nature and cross through each other, and simulations that have customers who miraculously disappear during the simulation run (this was not the programmers intention so it concerns verification, not validation).

Most simulation researchers agree that animation may be dangerous too, as the analysts and users tend to concentrate on very short simulation runs so the problems that occur only in long

runs go unnoticed. Of course, good analysts, who are aware of this danger, will continue the run long enough to create a rare event, which is then displayed to the users.

3. Validation

Once the analysts believe that the simulation model is programmed correctly, they must face the next question: is the conceptual simulation model (as opposed to the computer program) an accurate representation of the system under study (see Section 1)?

(A very old philosophical question is: do humans have accurate knowledge of reality or do they have only flickering images of reality, as Plato stated? In this paper, however, we take the view that managers act as if their knowledge of reality were sufficient. Also see Barlas and Carpenter (1990), Landry and Oral (1993), and Naylor, Balintfy, Burdick and Chu (1966, pp.310–320).)

This section discusses 1) obtaining real-world data, which may be scarce or abundant, 2) simple tests for comparing simulated and real data (namely graphical, Schruben–Turing, and t tests), 3) two new simple statistical procedures (based on regression analysis) for testing whether simulated and real responses are positively correlated and, possibly, have the same means too, 4) sensitivity analysis (using statistical design of experiments with its concomitant regression analysis) and risk analysis (based on Monte Carlo sampling), and 5) white and black box simulations.

3.1. Obtaining real-world data

System analysts must explicitly formulate the laws that they think govern the 'system under study', which is a system that already exists or is planned to be installed in the real world. The *system concept*, however, implies that the analysts must subjectively decide on the boundary of that system and on the attributes to be quantified in the model.

To obtain a valid model, the analysts should try to measure the inputs and outputs of the real system, and the attributes of intermediate variables. In practice, data are available in different quantities, as the next four situations illustrate.

1) Sometimes it is *difficult or impossible* to obtain relevant data. For example, in simulation studies of nuclear war, it is (fortunately) impossible to get the necessary data. In the simulation of whale population dynamics, a major problem is that data on whale behavior are hard to obtain. In the latter example more effort is needed for data collection. In the former example the analysts may try to show that the exact values of the input data are not critical. These problems will be further analyzed in the subsection on sensitivity analysis (Section 3.4.1).

2) Usually, however, it is possible to get *some* data. Typically the analysts have data only on the existing system variant or on a few historical variants; for example, the existing manufacturing system with its current scheduling rule.

3) In the military it is common to conduct *field tests* in order to obtain data on *future* variants. Kleijnen and Alink (1992) present a case study, namely mine hunting at sea by means of sonar: mine fields are created not by the enemy but by the friendly navy, and a mine hunt is executed in this field to collect data. Davis (1992a) and Fossett et al. (1991) also discuss several field tests for military simulations. Shannon (1975, pp.231–233) briefly discusses military field tests, too. Gray and Murray-Smith (1993) and Murray-Smith (1992) consider aeronautical field tests.

4) In some applications there is an *overload* of input data, namely if these data are collected electronically. For example, in the simulation of the performance of computer systems, the analysts use hardware and software monitors to collect data on the system state at regular time points (say, each nanosecond) or at each system state change (event). These data can be used to drive the simulation. Another example is provided by point-of-sale (POS) systems: based on the Universal Product Code (UPC) all transactions at the supermarket check-outs are recorded electronically (real-time data collection, data capture at the source); see Little (1991). In the near future more applications will be realized; for example, the geographical positions of trucks and

railroad cars will be determined and communicated electronically, and electronic data interchange (EDI) among companies will generate large quantities of data; see Geoffrion (1992) and Sussman (1992).

The further the analysts go back into the past, the more data they get and (as the next subsections will show) the more powerful the validation test will be, *unless* they go so far back that different laws governed the system. For example, many econometric models do not use data prior to 1945, because the economic infrastructure changed drastically during World War II. Of course, knowing when exactly different laws governed the system is itself a validation issue.

So real-world data may be either scarce or abundant. Moreover the data may show *observation error*, which complicates the comparison of real and simulated time series. Barlas (1989, p.72) and Kleijnen and Alink (1992) discuss observation errors in a theoretical and a practical situation respectively.

(The time series character of the model inputs and outputs, and the random noise are typical aspects of simulation. Other models – for example, inventory and econometric models – share some of these characteristics with simulation models. Validation of these other types of models does not seem to teach simulation analysts much.)

3.2. Some simple techniques for comparing simulated and real data

Suppose the analysts have succeeded in obtaining data on the real system (see the preceding subsection), and they wish to validate the simulation model. They should then feed real-world input data into the model, in *historical* order. In the simulation of computer systems this is called *trace driven* simulation. Davis (1992a, p.6) discusses the use of 'official data bases' to drive military simulations. After running the simulation program, the analysts obtain a time series of simulation output and compare that time series with the historical time series for the output of the existing system.

It is emphasized that in validation the analysts should not *sample* the simulation input from a (raw or smoothed) distribution of real-world input values. So they must use the historical input values in historical order. After they have validated the simulation model, they should compare different scenarios using sampled inputs, not historical inputs: it is 'certain' that history will never repeat itself exactly. As an example we consider a queueing simulation. To validate the simulation model, we use actual arrival times in historical order. Next we collect these arrival times in a frequency diagram, which we smooth formally by fitting an exponential distribution with a parameter (say) $\hat{\lambda}$. From this distribution we sample arrival times, using pseudorandom numbers. In sensitivity analysis we double the parameter $\hat{\lambda}$ to investigate its effect on the average waiting time.

Notice that validation of individual modules with *observable* inputs and outputs proceeds in exactly the same way as validation of the simulation model as a whole does. Modules with unobservable inputs and outputs can be subjected to sensitivity analyses (see Section 3.4.1).

How can system analysts compare a time series of simulation model output with a historical time series of real output? Several simple techniques are available:

1) The output data of the real system and the simulated system can be plotted such that the horizontal axis denotes time and the vertical axis denotes the real and simulated values respectively. The users may *eyeball timepaths* to decide whether the simulation model 'accurately' reflects the phenomena of interest. For example, do the simulation data in a business cycle study indicate an economic downturn at the time such a slump occurred in practice? Do the simulation data in a queueing study show the same saturation behavior (such as exploding queuelengths and blocking) as happened in the real system?

(Barlas (1989, p.68) gives a system dynamics example that seems to allow subjective graphical analysis only, since the time series (simulated and real) show 'highly transient, non-stationary behavior'.)

2) Another simple technique is the *Schruben–Turing* test. The analysts present a mixture of simulated and real time series to their clients,

and challenge them to identify (say) the data that were generated by computer. Of course, these clients may correctly identify some of the data by mere chance. This coincidence, however, the analysts can test statistically.

Turing introduced such an approach to validate Artificial Intelligence computer programs: users were challenged to identify which data (say, chess moves) were generated by computer, and which data were results of human reasoning. Schruben (1980) applies this approach to the validation of simulation models. He adds several statistical tests and presents some case studies. Also see Stanislaw (1986, p.182).

3) Instead of subjectively eyeballing the simulated and the real time series, the analysts can use *mathematical statistics* to obtain quantitative data about the quality of the simulation model. The problem, however, is that simulation output data form a time series, whereas practitioners are familiar with elementary statistical procedures that assume identically and independently distributed (i.i.d.) observations. Nevertheless it is easy to derive i.i.d observations in simulation (so that elementary statistical theory can be applied), as the next example will demonstrate.

Let $\underline{w}_i$ and $\underline{v}_i$ denote the *average* waiting time on day i in the simulation and the real system respectively. Suppose that n days are simulated and observed in reality respectively, so $i = 1, \ldots, n$. These averages, $\underline{w}_i$ and $\underline{v}_i$, do not need to be computed from a steady state time series of individual waiting times. They may be calculated from the individual waiting times of all customers arriving between 8 a.m. and 5 p.m. Then each day includes a start-up, transient phase. Obviously the simulated averages $\underline{w}_i$ are i.i.d. and so are the real averages $\underline{v}_i$. Suppose further that the historical arrival and service times are used to drive the simulation model. Statistically this trace-driven simulation means that there are n *paired* (correlated) differences $\underline{d}_i = \underline{w}_i - \underline{v}_i$, which are i.i.d. Then the t statistic analogous to (4) is

$$\underline{t}_{n-1} = \frac{\underline{\bar{d}} - \delta}{\underline{s}_d / \sqrt{n}}, \tag{5}$$

where $\underline{\bar{d}}$ denotes the average of the n $\underline{d}$'s, δ is the expected value of $\underline{d}$, and $\underline{s}_d$ represents the estimated standard deviation of $\underline{d}$.

(The variable $\underline{d}_i = \underline{w}_i - \underline{v}_i$ denotes the difference between simulated and real average waiting time on day i when using the same arrival and service times. Hence $\underline{\bar{d}}$ is the average of the n differences between the n average simulated and n average real waiting times per day. Other statistics of interest may be the percentage of customers waiting longer than (say) one minute, the waiting time exceeded by only 10% of the customers, etc. Testing these statistics is discussed in Kleijnen and Van Groenendaal (1992, pp.195–197).)

Suppose that the null-hypothesis is H_0: $\delta = 0$, and (5) gives a value $\underline{t}_{n-1}$ that is significant ($|\underline{t}_{n-1}| > t_{n-1;\alpha/2}$). Then the simulation model is rejected, since this model gives average waiting times per day that deviate significantly from reality. In case of a non-significant $|\underline{t}_{n-1}|$ the conclusion is that the simulated and the real means are 'practically' the same so the simulation is 'valid enough'. This interpretation, however, deserves some comments.

Strictly speaking, the simulation is only a model, so δ (the expected value of $\underline{d}$ and hence the expected value of $\underline{\bar{d}}$) is never exactly zero. Let us consider three points.

1) The bigger the sample size is, the smaller the critical value $t_{n-1;\alpha/2}$ is; for example, for a fixed $\alpha = 0.05$ but $n = 5$ and 121 respectively, $t_{n-1;\alpha/2} = 2.776$ and 1.980 respectively. So, all other things being equal, a simulation model has a higher chance of being rejected as its sample size is bigger.

2) Simulating 'many' days ('large' n) gives a 'precise' estimate $\underline{\bar{d}}$ and hence a significant $\underline{t}_{n-1}$ (in Eq.(5), $\underline{s}_d / \sqrt{n}$ goes to zero because of n; in the numerator, $\underline{\bar{d}}$ has expected value different from 0; so the test statistic $\underline{t}_{n-1}$ goes to infinity, whereas the critical value $t_{n-1;\alpha/2}$ goes to $z_{\alpha/2}$, which denotes the $1 - \alpha/2$ quantile of the standard normal variable). So model mis-specification would always lead to rejection if the sample size n were infinite.

3) The t statistic may be significant and yet unimportant. If the sample is very large, then the t statistic is nearly always significant for $\delta \neq 0$;

nevertheless the simulated and the real means may be 'practically' the same so the simulation is 'valid enough'. For example, if $E(\underline{w}_i) = 1000$ and $E(\underline{v}_i) = 1001$ (so $\delta = 1$), then the simulation model is good enough for all practical purposes. Also see Fleming and Schoemaker (1992, p.472).

In general, when testing the validity of a model through statistics such as (5), the analysts can make either a 'type I' or a 'type II' error. So they may reject the model while the model is valid: type I or α error. Or they may accept the model while the model is not valid: type II or β error. The probability of a β error is the complement of the 'power' of the test, which is the probability of rejecting the model when the model is wrong indeed. The probability of a type I error in simulation is also called the *model builder's risk*; the type II error probability is the *model user's risk*.

The power of the test of H_0: $\delta = 0$ increases as the model specification error (the 'true' δ) increases. For example, as (the true) δ goes to infinity so does $\underline{t}_{n-1}$ in (5), hence the simulation model is rejected (for any n and α, which fix $t_{n-1;\alpha/2}$). (This power can be computed through the 'non-central' t statistic, which is a t statistic with non-zero mean.) A significance or 'critical' level α (used in $t_{n-1;\alpha/2}$) means that the type I error probability equals α. The probability of a β error increases as α decreases, given a fixed number of simulated days: as α decreases, the critical value $t_{n-1;\alpha/2}$ increases. To keep the type I probability fixed and to decrease the type II probability, the analysts may increase the number of simulated days: if α is kept constant and n increases, then $t_{n-1;\alpha/2}$ decreases.

The analysts may also make the t test more powerful by applying *variance reduction techniques* (VRTs), such as control variates (see Eq. (1)). If control variates work, they decrease the variance of $\underline{w}$ and hence the variance of $\underline{d}$ $(= \underline{w} - \underline{v})$. Then $\underline{s}_d$ in (5) has a smaller expected value, and the probability of a high $\underline{t}_{n-1}$ increases. The simplest and most popular VRT is common (pseudo)random numbers. Running the simulation with real-world inputs is a form of this VRT. It decreases var($\underline{d}$) (not var($\underline{w}$)).

Balci and Sargent (1984b) analyze the theoretical tradeoffs among α and β error probabilities, sample size, and so on.

The selection of a value for α is problematic. Popular values are 0.10 and 0.05. Theoretically, the analysts should determine these values by accounting for the financial consequences – or more generally, the disutilities – of making type I and type II errors respectively. Such an approach is indeed followed in decision theory and in Bayesian analysis; see Bodily (1992), Kleijnen (1980, pp.115–134) and also Davis (1992a, p.20). Because the quantification of these utility functions is extremely difficult in most simulation studies, this paper follows classic statistical theory.

3.3. Two new simple statistical tests for comparing simulated and real data

Two tests based on new interpretations of classic tests in regression analysis are discussed in this subsection.

1) Consider again the example where $\underline{w}_i$ and $\underline{v}_i$ denoted the average waiting time on day i in the simulation and the real system respectively, which use the same inputs. Suppose that on day 4 the real average waiting time is relatively high, that is, higher than expected (because service times were relatively high on that day): $v_4 > E(\underline{v})$. Then it seems reasonable to require that on that day the simulated average (which uses the same service times) is also relatively high: $w_4 > E(\underline{w})$. So the new test checks that $\underline{v}$ and $\underline{w}$ are *positively correlated*: H_0: $\rho > 0$ where ρ denotes their linear correlation coefficient. (They *might* have the same mean so $\delta = 0$ in Eq. (5).) So the analysts may then formulate a *less stringent* validation test: simulated and real responses do not necessarily have the same mean, but they are positively correlated.

To investigate this correlation, the analysts may plot the n pairs (v_i, w_i). That graphical approach can be formalized through the use of the *ordinary least squares* (OLS) algorithm. Testing the hypothesis of positively correlated $\underline{v}$ and $\underline{w}$ is simple if $\underline{v}$ and $\underline{w}$ are *bivariate normally* distributed. This is a realistic assumption in the example,

because of a central limit theorem (see the comment on Eq. (4)). It can be proved that such a bivariate normal distribution implies a linear relationship between the conditional mean of one variable and the value of the other variable:

$$E(\underline{w} \mid \underline{v} = v) = \beta_0 + \beta_1 v. \quad (6)$$

So the analysts can use OLS to estimate the intercept and slope of the straight line that passes through the 'cloud' of points (v_i, w_i). The proposed test concerns the one-sided hypothesis

$$H_0 : \beta_1 \leq 0. \quad (7)$$

To test this null-hypothesis, a t statistic can be used, as any textbook on regression analysis shows. This test means that the analysts reject the null-hypothesis and accept the simulation model if there is strong evidence that the simulated and the real responses are *positively* correlated.

2) Sometimes simulation is meant to predict *absolute responses* (not relative responses corresponding to different scenarios; for example, what is the effect of adding one server to a queueing system?). For example, in the mine hunting case study (Kleijnen and Alink, 1992) one of the questions concerns the probability of detecting mines in a certain area: is that probability so high that it makes sense to do a mine sweep? The analysts may then formulate a *more stringent* test:

(i) the means of $\underline{w}$ (the simulated response) and $\underline{v}$ (the historical response) are identical, *and*

(ii) if a historical observation exceeds its mean, then the corresponding simulated observation tends to exceed its mean too.

These two conditions lead to the *composite hypothesis*

$$H_0 : \beta_0 = 0 \text{ and } \beta_1 = 1, \quad (8)$$

which implies $E(\underline{w}) = E(\underline{v})$ (which was also tested through Eq. (5)) and is more specific than Eq. (7) is.

(Note that $\beta_1 = \rho\sigma_w/\sigma_v$. So if $\beta_1 = 1$ and $\rho < 1$ then $\sigma_w > \sigma_v$: if the model is not perfect ($\rho < 1$), then its variance exceeds the real variance.)

To test this composite hypothesis, the analysts should compute the Sum of Squared Errors (SSE) with and without that hypothesis (which correspond with the 'reduced' and the 'full' regression model respectively), and compare these two values. If the resulting F statistic is significantly high, the analysts should reject the hypothesis and conclude that the simulation model is not valid. Details on this F test can be found in Kleijnen and Van Groenendaal (1992, pp.209–210).

Statistical tests require many observations to make them powerful. In validation however, there are often not many observations on the real system (see Section 3.1). Sometimes, however, there are very many observations. Then not only the means of the simulated and the real time series and their (cross)correlation ρ can be compared, but also the autocorrelations corresponding with lag 1, 2, etc. Spectral analysis is a sophisticated technique that estimates the autocorrelation structure of the simulated and the historical time series respectively, and compares these two structures. Unfortunately, that analysis is rather difficult (and – as stated – requires long time series). Barlas (1989, p.61) criticizes Box–Jenkins models for the same reasons.

Note that Fleming and Schoemaker (1992) discuss the use of regression plots in case of multiple outputs.

3.4. Sensitivity analysis and risk analysis

3.4.1. Sensitivity analysis

Models and submodels (modules) with *unobservable* inputs and outputs can not be subjected to the tests of Section 3.2 and Section 3.3. The analysts should then apply sensitivity analysis, in order to determine whether the model's behavior agrees with the judgments of the experts (users and analysts). In case of observable inputs and outputs sensitivity analysis is also useful, as this subsection will show. (The observability of systems is also discussed in Zeigler (1976).)

Sensitivity analysis or *what-if analysis* is defined in this paper as the systematic investigation of the reaction of model outputs to drastic changes in model inputs and model structure: global (not local) sensitivities. For example, what are the effects if in a queueing simulation the arrival rate doubles; what if the priority rule changes from FIFO to LIFO?

The techniques for sensitivity analysis discussed in this paper, are design of experiments and regression analysis. Unfortunately, most practitioners apply an inferior *design of experiments*: they change one simulation input at a time. Compared with (fractional) factorial designs (such as 2^{K-P} designs), the 'one at a time' designs give estimated effects of input changes that have higher variances (less accurate). Moreover, these designs cannot estimate interactions among inputs. See Kleijnen and Van Groenendaal (1992, pp.167–179).

How can the results of experiments with simulation models be analyzed and used for interpolation and extrapolation? Practitioners often *plot* the simulation output (say) y versus the simulation input x_k, one plot for each input k with $k = 1, \ldots, K$. (For example, if the arrival and service rates are changed in an M/M/1 simulation then $K = 2$.) More refined plots are conceivable, for instance, superimposed plots. Also see the 'spiderplots' and 'tornado diagrams' in Eschenbach (1992).

This practice can be formalized through *regression analysis*. So let y_i denote the simulation response (for example, average waiting time per day) in combination (or run) i of the K simulation inputs, with $i = 1, \ldots, n$, where n denotes the total number of simulation runs. Further let x_{ik} be the value of simulation input k in combination i, β_k the main or first order effect of input k, $\beta_{kk'}$ the interaction between inputs k and k', and e_i the approximation (fitting) error in run i. Then the input/output behavior of the simulation model may be approximated through the regression (meta)model

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \beta_{kk'} x_{ik} x_{ik'} + e_i. \tag{9}$$

Of course, the validity of this approximation must be tested. *Cross-validation* uses some simulation inputs and the concomitant output data to get estimated regression parameters $\hat{\beta}$. Next it employs the estimated regression model to compute the forecast $\hat{y}$ for some other input combinations. The comparison of forecasted output $\hat{y}$ and simulated output y is used to validate the regression model. See Kleijnen and Van Groenendaal (1992, pp.156–157).

Inputs may be *qualitative*. An example is the priority rule in a queueing simulation. Technically, binary variables (x_{ik} is zero or one) are then needed; see Kleijnen (1987).

An example of experimental design and regression analysis is provided by Kleijnen, Rotmans and Van Ham (1992). They apply these techniques to several modules of a (deterministic) simulation model of the greenhouse effect of carbon dioxide (CO_2) and other gases. This approach gives estimates $\hat{\beta}$ of the effects of the various inputs. These estimated effects should have the right *signs*: the users (not the statisticians) know that certain inputs increase the global temperature. Wrong signs indicate computer errors (see Section 2) or conceptual errors. Indeed Kleijnen et al. (1992, p.415) give examples of sensitivity estimates with the wrong signs, which lead to correction of the simulation model. One more example is given by Kleijnen and Alink (1992). The role of experimental design in V&V of simulation models is also discussed in Gray and Murray-Smith (1993), Murray-Smith (1992), and Pacheco (1988).

Classic experimental designs (with $n > K$), however, require too much computer time, when the simulation study is still in its early (pilot) phase. Then *very many inputs* may be conceivably important. Bettonvil and Kleijnen (1991) derive a *screening* technique based on sequential experimentation with the simulation model. They split up (bifurcate) the aggregated inputs as the experiment proceeds, until finally the important individual inputs are identified and their effects are estimated. They apply this technique to the ecological simulation mentioned above. In this application there are 281 inputs. It is remarkable that this statistical technique identifies some inputs that were originally thought to be unimportant by the users.

The *magnitudes* of the sensitivity estimates show which inputs are important. For important inputs the analysts should try to collect data on the input values that may occur in practice. If the

analysts succeed, then the validation techniques of the preceding subsections can be applied.

(If the simulation inputs are under the decision makers' control, then these inputs should be steered in the right direction. The regression (meta)model can help the analysts determine the directions in which those inputs should be steered. For example, in the greenhouse case the governments should restrict emissions of the gases concerned.)

Before executing the experimental design (either a one at a time or a fractional factorial design), the analysts must determine the experimental domain or experimental frame. The design tells *how* to explore this domain, using the expertise of the statistician. Zeigler (1976, p.30) defines the *experimental frame* as "a limited set of circumstances under which the real system is to be observed or experimented with". He emphasizes that "a model may be valid in one experimental frame but invalid in another". This paper (Section 3.1) has already mentioned that going far back into the past may yield historical data that are not representative of the current system; that is, the old system was ruled by different laws. Similarly, a model is accurate only if the values of its input data remain within a certain area. For example, Bettonvil and Kleijnen's (1991) screening study shows that the greenhouse simulation is valid, only if the simulation input values range over a relatively small area. Some authors (for example, Banks, 1989, and Barlas, 1989), however, claim that a model should remain valid under *extreme* conditions. This paper rejects that claim, but perhaps this disagreement is a matter of definition: what is 'extreme'?

So the simulation model is valid within a certain area of its inputs only (the area may be defined as the K-dimensional hypercube formed by the K input ranges). Within that area the simulation model's input/output behavior may vary. For example, a *first* order regression (meta)model (see Eq.(9) with the double summation term eliminated) is a good approximation of the input/output behavior of a simulated M/M/1 system, only if the traffic load is 'low'. When traffic is heavy, a second order regression model or a logarithmic transformation may apply.

Our conclusion is that sensitivity analysis should be applied to find out which inputs are really important. That information is useful, even if there are many data on the input and output of the simulated system (see the first paragraph of Section 3.4.1). Collecting information on the important inputs – if possible – is worth the effort. However, it may be impossible or impractical to collect reliable information on those inputs, as the examples of the whale and the nuclear attack simulations have already demonstrated (see Section 3.1). Then the analysts may apply the following technique.

3.4.2. Risk analysis

In *risk analysis* or *uncertainty analysis* the analysts first derive a probability distribution of input values, using the clients' expert knowledge. Next they use Monte Carlo sampling to generate input values from those distributions. These values are fed into the simulation model, which yields a probability distribution of output values. Technical details and applications are given by Bodily (1992), Kleijnen and Van Groenendaal (1992, pp.75–78), and Krumm and Rolle (1992).

The study of the sensitivity to the input distributions assumed in the risk analysis may be called *robustness analysis*. The relationships among sensitivity, risk, and robustness analyses require more research; see Kleijnen (1994).

3.5. White box simulation versus black box simulation

Karplus (1983) perceives a whole spectrum of mathematical models (not only simulation models), ranging from *black box* (noncausal) models in the social sciences through gray box models in ecology to *white box* (causal) models in physics and astronomy. What does this classification scheme mean for the validation of simulation models, especially in operations research (OR)?

(This range of model types is also found in OR: examples are regression analysis (black box), linear programming (gray box), and inventory control (white box). Also see Oral and Kettani (1993).)

A typical aspect of many simulation studies is

that their *conceptual* models are based on common sense and on direct observation of the real system: *white box* simulation. For example, logistic problems in a factory may be studied through a simulation program that models the factory as a queueing network. This model can directly incorporate intuitive knowledge about the real system: a job arrives, looks for an idle machine in the first stage of the production process, leaves the machine upon completion of the required service, goes to the second stage of its fabrication sequence, and so on (if expediting of jobs is observed in the real system, then this complication can be included in the simulation). Counter-intuitive behavior of the model may indicate either programming and modeling errors or new insight (surprise value of information; see Kleijnen, 1980, pp.115–134, and Richardson and Pugh, 1981, pp.317–319).

The analysts can further apply a *bottom-up* approach: connecting the submodels (or modules) for the individual factory departments, they develop the total simulation model. In this way the simulation grows in complexity and – hopefully – realism. (Davis (1992b) examines combining models of different resolution (aggregation) that were not originally designed to be combined. Bankes (1993) criticizes large simulation models used in policy analysis.)

Animation is a good means to obtain *face validity* of white box simulation models. Moreover, many white box systems have relatively many data available (so Karplus's classification is related, not orthogonal, to the classification used in this paper). Then the statistical tests discussed in Section 3.2 and Section 3.3 can be applied.

In some application areas, however, simulation models are *black box* models. Examples are plentiful in aggregated econometric modeling: macroeconomic consumption functions relate total national consumption to Gross National Product (GNP); see Kleijnen and Van Groenendaal (1992, pp.57–69). The validation of black box models is more difficult, since (by definition) the analysts can not measure the internal relationships and the internal data of these models. Maybe they can measure input and output data, and apply the tests of Section 3.2 and Section 3.3; also see Bankes (1993) and Pagan (1989). Models and submodels with *unobservable* inputs and outputs can be subjected to the sensitivity analysis of Section 3.4.1.

In black box models the emphasis in validation is on *prediction*, not *explanation*. Nevertheless sensitivity analysis of black box models may give estimated effects of various inputs that have wrong signs. These wrong signs indicate computer errors or conceptual errors. Prediction versus explanation in validation is discussed in more detail in Davis (1992a, pp.7–10).

Some analysts use model *calibration*, that is, they adjust the simulation model's parameters (using some minimization algorithm) such that the simulated output deviates minimally from the real output. (Obviously, those latter data can *not* be used to validate the model.) Examples can be found in ecological modeling; see Beck (1987). Another example is provided by the mine hunting simulation in Kleijnen and Alink (1992), which uses an artificial parameter to steer the simulation response into the direction of the observed real responses. Calibration is a last resort employed in black box simulation. Davis (1992b) discusses how aggregated models can be calibrated using detailed models. Also see Bankes (1993, p.443).

4. Documentation, assessment, credibility, and accreditation

The model's assumptions and input values determine whether the model is valid, *and will remain valid* when the real system and its environment will change: model maintenance problem. Therefore the analysts should provide information on these assumptions and input values in the model's documentation. In practice, however, many assumptions are left implicit, deliberately or accidently. And input data including scenarios are left undocumented. (Davis (1992a, p.4) distinguishes between 'bare model' and 'data base', which corresponds with the terms 'model' and 'input data' in this paper.)

V&V are important components of *assessment*, defined as "a process by which interested

parties (who were not involved in a model's origins, development, and implementation) can determine, with some level of confidence, whether or not a model's result can be used in decision making" (Fossett et al., 1991, p.711). To enable users to assess a simulation model, it is necessary to have good documentation. Assessment is discussed at length in Davis (1992a); also see Oral and Kettani (1993, p.229).

Credibility is "the level of confidence in [a simulation's] results"; see Fossett et al. (1991, p.712). These authors present a framework for assessing this credibility. That framework comprises 14 inputs. These inputs have also been discussed in this paper, explicitly or implicitly. They apply their framework to three military weapon simulations.

V&V are important components of *accreditation*, which is "an official determination that a model is acceptable for a specific purpose", see Davis (1992a), Gass (1993), and Williams and Sikora (1991).

The present paper shows that V&V have many aspects, involve different parties, and require good documentation. Gass (1984) proposes to produce four manuals, namely for analysts, users, programmers, and managers respectively.

(The lack of good documentation is a problem, not only with simulation programs but also with other types of mathematical models and with software in general; see Section 2.1.)

5. Supplementary literature

V&V of simulation models have been discussed in many textbooks on simulation. Examples are Banks and Carson (1984), Law and Kelton (1991, pp.298–324), and Pegden et al. (1990, pp.133–162). These books give many additional references Stanislaw (1986) gives many references to the behavioral sciences.

Some case studies were mentioned above. In addition, Kleijnen (1993) gives a production-planning case study, Carson (1989) presents a cigarette fabrication study, and Davis (1992a) gives summaries of several military studies.

Dekker, Groenendijk and Sliggers (1990) discuss V&V of models that are used to compute air pollution. These models are employed to issue permits for building new factories and the like.

Banks (1989) proposes control charts, which are well-known from quality control. Reckhow (1989) discusses several more statistical techniques.

Hodges (1991) gives a more polemical discussion of validation.

Findler and Mazur (1990) present an approach based on Artificial Intelligence methodology, to verify and validate simulation models.

In case no data are available, Diener, Hicks and Long (1992) propose to compare the new simulation model to the old well-accepted but non-validated simulation model, assuming the latter type of simulation is available. Also see Murray-Smith (1992).

Balci and Sargent (1984a) and Youngblood (1993) give detailed bibliographies. The references of this paper augment those bibliographies.

6. Conclusions

This paper surveyed verification and validation (V&V) of models, especially simulation models in operations research. It emphasized statistical techniques that yield reproducible, objective, quantitative data about the quality of simulation models.

For *verification* it discussed the following techniques (see Section 2):

1) general good programming practice such as modular programming;

2) checking of *intermediate* simulation outputs through tracing and statistical testing per module (for example, the module for sampling random variables);

3) comparing *final* simulation outputs with analytical results for simplified simulation models, using statistical tests;

4) animation.

For *validation* it discussed the following techniques (see Section 3):

1) obtaining real-world data, which may be scarce or abundant;

2) simple tests for comparing simulated and real data: graphical, Schruben-Turing, and t tests;

3) two new simple statistical procedures for testing whether simulated and real responses are positively correlated and, possibly, have the same means too;

4) sensitivity analysis (based on design of experiments and regression analysis) and risk analysis (Monte Carlo sampling) for estimating which inputs are really important and for quantifying the risks associated with inputs for which no data can be obtained at all, respectively;

5) white and black box simulations.

Both verification and validation require good documentation. V&V are crucial parts of assessment, credibility, and accreditation. Supplementary literature on V&V is given for further study.

This essay demonstrates the usefulness of mathematical statistics in V&V. Nevertheless, analysts and users of a simulation model should be convinced of its validity, not only by statistics but also by other procedures; for example, animation (which may yield face validity).

It seems impossible to prescribe a fixed order for applying the various V&V techniques. In some applications certain techniques do not apply at all. Practice shows that V&V techniques are applied in a haphazard way. Hopefully, this paper stimulates simulation analysts and users to pay more attention to the various aspects of V&V and to apply some of the techniques presented in this paper. The taxonomy discussed in this paper in detail, and the other taxonomies referred to, may also serve as checklists for practitioners. Nevertheless, simulation will remain both an art as well as a science.

Acknowledgements

The reviews by three referees lead to drastic reorganizations and expansions of previous versions of this paper.

References

Adrion, W.R., Branstad, M.A., and Cherniavsky, J.C. (1982), "Validation, verification and testing of computer software", *ACM Computing Surveys* 14, 159–192.

Baber, R. (1987), *The Spine of Software; Designing Provable Correct Software: Theory and Practice*, Wiley, Chichester.

Balci, O., and Sargent, R.G. (1984a), "A bibliography on the credibility, assessment and validation of simulation and mathematical models," *Simuletter* 15 3, 15–27.

Balci, O., and Sargent, R.G. (1984b), "Validation of simulation models via simultaneous confidence intervals," *American Journal of Mathematical and Management Science* 4 3–4, 375–406.

Bankes, S. (1993), "Exploratory modeling for policy analysis," *Operations Research* 41 3, 435–449.

Banks, J. (1989), "Testing, understanding and validating complex simulation models," in: *Proceedings of the 1989 Winter Simulation Conference*.

Banks, J., and Carson, J.S. (1984), *Discrete-event System Simulation*, Prentice-Hall, Englewood Cliffs, NY.

Barlas, Y. (1989), "Multiple tests for validation of system dynamics type of simulation models," *European Journal of Operational Research* 42 1, 59–87.

Barlas, Y. (1990), "An autocorrelation function test for output validation," *Simulation* 56, 7–16.

Barlas, Y., and Carpenter, S. (1990), "Philosophical roots of model validation: Two paradigms," *System Dynamics Review* 6 2, 148–166.

Beck, M.B. (1987), "Water quality modeling: A review of the analysis of uncertainty," *Water Resources Research* 23 8, 1393–1442.

Benbasat, I., and Dhaliwal, J.S. (1989), "A framework for the validation of knowledge acquisition," *Knowledge Acquisition* 1, 215–233.

Bettonvil, B., and Kleijnen, J.P.C. (1991), "Identifying the important factors in simulation models with many factors," Tilburg University.

Bodily, S.E. (1992), "Introduction; the practice of decision and risk analysis," *Interfaces* 22 6, 1–4.

Carson, J.S. (1989), "Verification and validation: A consultant's perspective," in: *Proceedings of the 1989 Winter Simulation Conference*.

Dahl, O. (1992), *Verifiable Programming*, Prentice-Hall, Englewood Cliffs, NY.

Davis, P.K. (1992a), "Generalizing concepts of verification, validation, and accreditation (VV&A) for military simulation," RAND, October 1992a (to be published as R-4249-ACQ).

Davis, P.K. (1992b), "An introduction to variable-resolution modeling and cross-resolution model connection," RAND, October 1992b (to be published as R-4252-DARPA).

Dekker, C.M., Groenendijk, A., and Sliggers, C.J. (1990), "Kwaliteitskriteria voor modellen om luchtverontreiniging te berekenen" (Quality criteria for models to compute air pollution), Report 90, VROM, Leidschendam, Netherlands.

DeMillo, R.A., McCracken, W.M., Martin, R.J., and Passafiume, J.F. (1987), *Software Testing and Evaluation*, Benjamin/Cummings, Menlo Park, CA.

Diener, D.A. Hicks, H.R., and Long, L.L. (1992), "Comparison of models: Ex post facto validation/acceptance?," in: *Proceedings of the 1992 Winter Simulation Conference.*

Eschenbach, T.G. (1992), "Spiderplots versus tornado diagrams for sensitivity analysis", *Interfaces* 22/6, 40–46.

Findler, N.V., and Mazur, N.M. (1990), "A system for automatic model verification and validation", *Transactions of the Society for Computer Simulation* 6/3, 153–172.

Fleming, R.A., and Schoemaker, C.A. (1992), "Evaluating models for spruce budworm-forest management: Comparing output with regional field data", *Ecological Applications* 2/4, 466–477.

Fossett, C.A., Harrison, D., Weintrob, H., and Gass, S.I. (1991), "An assessment procedure for simulation models: A case study", *Operations Research* 39/5, 710–723.

Gass, S.I. (1984), "Documenting a computer-based model", *Interfaces* 14/3, 84–93.

Gass, S.I. (1993), "Model accreditation: A rationale and process for determining a numerical rating", *European Journal of Operational Research* 66/2, 250–258.

Geoffrion, A.M. (1992), "Forces, trends and opportunities in MS/OR", *Operations Research* 40/3, 423–445.

Gray, G.J. and Murray-Smith, D.J. (1993), "The external validation of nonlinear models for helicopter dynamics", in: R. Pooley and R. Zobel (eds.), *Proceedings of the United Kingdom Simulation Society Conference*, UKSS.

Hodges, J.S. (1991), "Six (or so) things you can do with a bad model", *Operations Research* 39/3, 355–365.

Karplus, W.J. (1983), "The spectrum of mathematical models", *Perspectives in Computing* 3/2, 4–13.

Kleijnen, J.P.C. (1974), *Statistical Techniques in Simulation, Part I*, Marcel Dekker, New York.

Kleijnen, J.P.C. (1980), *Computers and Profits: Quantifying Financial Benefits of Information*, Addison-Wesley, Reading, MA.

Kleijnen, J.P.C. (1987), *Statistical Tools for Simulation Practitioners*, Marcel Dekker, New York.

Kleijnen, J.P.C. (1993), "Simulation and optimization in production planning: A case study", *Decision Support Systems* 9, 269–280.

Kleijnen, J.P.C. (1994), "Sensitivity analysis versus uncertainty analysis: When to use what?", in: *Proceedings Predictability and Nonlinear Modeling in Natural Sciences and Economics*, Kluwer, Dordrecht.

Kleijnen, J.P.C., and Alink, G.A. (1992), "Validation of simulation models: Mine-hunting case study", Tilburg University.

Kleijnen, J.P.C., and Van Groenendaal, W. (1992), *Simulation: A Statistical Perspective*, Wiley, Chichester.

Kleijnen, J.P.C., Rotmans, J., and Van Ham, G. (1992), "Techniques for sensitivity analysis of simulation models: A case study of the CO_2 greenhouse effect", *Simulation* 58/6, 410–417.

Knepell, P.L., and Arangno, D.C. (1993), *Simulation validation: A confidence assessment methodology*, IEEE Computer Society Press, Los Alamitos, CA.

Krumm, F.V., and Rolle, C.F. (1992), "Management and application of decision and risk analysis in Du Pont", *Interfaces* 22/6, 84–93.

Landry, M., and Oral, M. (1993), "In search of a valid view of model validation for operations research", *European Journal of Operational Research* 66/2, 161–167.

Law, A.M., and Kelton, W.D. (1991), *Simulation Modeling and Analysis, 2nd ed.*, McGraw-Hill, New York.

Little, J.D.C. (1991), "Operations research in industry: New opportunities in a changing world", *Operations Research* 39/4, 531–542.

Majone, G., and Quade, E.S. (1980), *Pitfalls of Analysis*, Wiley, Chichester.

Miser, H.J. (1993), "A foundational concept of science appropriate for validation in operational research", *European Journal of Operational Research* 66/2, 204–215.

Murray-Smith, D.J. (1992), "Problems and prospects in the validation of dynamic models" in: A. Sydow (ed.), *Computational Systems Analysis 1992*, Elsevier, Amsterdam.

Naylor, T.H., Balintfy, J.L., Burdick, D.S., and Chu, K. (1966), *Computer Simulation Techniques*, Wiley, New York.

Oral, M., and Kettani, O. (1993), "The facets of the modeling and validation process in operations research", *European Journal of Operational Research* 66/2, 216–234.

Pace, D.K. (1993), "A paradigm for modern modeling and simulation, verification, validation and accreditation", Johns Hopkins University, Laurel, MD.

Pacheco, N.S. (1988), "Session III: Simulation certification, verification and validation", in: *SDI Testing: the Road to Success; 1988 Symposium Proceedings International Test & Evaluation Association*, ITEA, Fairfax, VA, 22033.

Pagan, A. (1989), "On the role of simulation in the statistical evaluation of econometric models", *Journal of Econometrics* 40, 125–139.

Pegden, C.P., Shannon, R.E., and Sadowski, R.P. (1990), *Introduction to Simulation using SIMAN*, McGraw-Hill, New York.

Reckhow, K.H. (1989), "Validation of simulation models: Philosophical and statistical methods of confirmation", in: M.D. Singh (ed.), *Systems & Control Encyclopedia*, Pergamon Press, Oxford.

Richardson, G.H., and Pugh, A. (1981), *Introduction to System Dynamics Modeling with DYNAMO*, MIT Press, Cambridge, MA.

Ripley, B.D., "Uses and abuses of statistical simulation", *Mathematical Programming* 42, 53–68.

Sargent, R.G. (1991), "Simulation model verification and validation", in: *Proceedings of the 1991 Winter Simulation Conference*.

Schriber, T.J. (1991), *An Introduction to Simulation Using GPSS/H*, Wiley, New York.

Schruben, L.W. (1980), "Establishing the credibility of simulations", *Simulation* 34/3, 101–105.

Shannon, R.E. (1975), *Systems Simulation: The Art and Science*, Prentice-Hall, Englewood Cliffs, NJ.

Stanislaw, H. (1986), "Tests of computer simulation validity: What do they measure?", *Simulation & Games* 17/2, 173–191.

Sussman, J.M. (1992), "Intelligent vehicle highway systems: A challenge awaits the transportation and OR/MS community", *OR/MS Today* 19/6, 18–23.

Tocher, K.D. (1963), *The Art of Simulation*, English University Press, London.

Wang W., Yin, H., Tang, Y., and Xu, Y. (1993), "A methodology for validation of system and sub-system level models", Department of System Engineering and Mathematics, National University of Defense Technology, Changsha, Hunan, 410073, P.R. China.

Whitner, R.B., and Balci, O. (1989), "Guidelines for selecting and using simulation model verification techniques", in: *Proceedings of the 1989 Winter Simulation Conference*.

Williams, M.K., and Sikora, J. (1991), "SIMVAL Minisymposium – A report", *Phalanx, The Bulletin of Military Operations Research*, 24/2, **PAGES?**

Youngblood, S.M. (1993), "Literature review and commentary on the verification, validation and accreditation of models and simulations", Johns Hopkins University, Laurel, MD.

Zeigler, B. (1976), *Theory of Modelling and Simulation*, Wiley Interscience, New York.