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

VALIDATION OF TRACE-DRIVEN SIMULATION MODELS: REGRESSION ANALYSIS REVISITED

Jack P. C. Kleijnen
Bert Bettonvil
Willem Van Groenendahl

Department of Information Systems and Auditing (BIKA)/Center for Economic Research (CentER)
School of Management and Economics (FEW)
Tilburg University (KUB)
5000 LE Tilburg, NETHERLANDS

ABSTRACT

For the validation of trace-driven simulation models this paper recommends a simple statistical test that uses elementary regression analysis in a novel way. This test concerns a (joint) null-hypothesis: the outputs of the simulated and the real systems have the same means and the same variances. Technically, the differences between simulated and real outputs are regressed on their sums, and the resulting slope and intercept are tested to be zero. This paper further proves that it is wrong to use a naive test that regresses the simulation outputs on the real outcomes, and hypothesizes that the resulting regression line gives a 45° line through the origin. The new and the old tests are investigated in Monte Carlo experiments with inventory systems. The conclusion is that the new test has the correct type I error probability, whereas the old test (falsely) rejects a valid simulation model substantially more often than the nominal alpha level. The power of the new test increases, as the simulation model deviates more from the real system.

1 INTRODUCTION

This paper is the companion paper of Kleijnen, Bettonvil, and Van Groenendaal (1996), which has been accepted (conditionally) by *Management Science*. Both papers concern a novel test for the validation of trace-driven simulations. The *Management Science* paper estimates the statistical performance of that test, using a Monte Carlo study of single-server queueing simulations (namely, M/G/1), whereas this paper illustrates that performance through single-item inventory simulations (see §3 for details). Moreover, because of page restrictions Kleijnen, et al. (1996) covers only parts of the original working paper; this paper includes other parts of that working paper, and adds recent references.

The remainder of this section answers the following questions:
(i) What is meant by validation?
(ii) What has the literature to say about validation?
(iii) What is the contribution of this paper?
(iv) How is this paper organized?
Hasty readers may skip the next two subsections (§1.1 and 1.2).

1.1 Definition of Validation

This paper uses the following definition in the classic textbook by Law and Kelton (1991, p. 299): '*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'. To illustrate some validation issues, consider the following practical problem.

The management of a inventory system wants to control the total costs of their system, which consists of stock-carrying costs, ordering costs, and lost-sales costs (no backordering). To solve this problem, the Management Science/Operations Research (MS/OR) specialists build a simulation model that represents this inventory system. Before using that model to advise management, the MS/OR experts wish to validate their model; that is, determine whether the model is an accurate representation of the real inventory system.

Obviously, validation should not aim at a *perfect* model: the perfect model would be the real system itself. So, validation is interpreted here as comparing *data* on the real and the simulated systems.

Those data pertain to *inputs* and *outputs*; for example, customer demand per day and order lead times (which are stochastic inputs) and total inventory costs per day (which measures the output).

Comparing the output data of the real and simulated systems makes more sense if both systems are observed under *similar circumstances*: the analysts should not compare total costs during a period that includes a *long* lead time in the real system with the costs during a simulated period with a *short* lead time: the former period has more lost sales than the latter period has.

Hence, for validation purposes the analysts should feed real-world input data into the model, in *historical* order (assuming such data are available indeed). This is called *trace driven* simulation in computer performance modeling; we shall use this term throughout this paper. Law and Kelton (1991, p. 316) call this the 'correlated inspection approach' (after this validation phase, 'production runs' will follow). After running the simulation program, the analysts obtain simulation output; they compare that output with the historical output of the existing system.

Note: After this trace-driven validation, the analysts should use the historical input data to develop a (sub) model for the input. For example, they may specify a particular type of distribution (say, the Gaussian distribution) for the demand variable, possibly incorporating autocorrelations and time trends. After estimation of the parameters of that distribution, they may apply goodness-of-fit tests to verify whether this distribution gives an adequate approximation of this input. See Kleijnen (1974, pp. 68-69).

Note: Regression analysis of trace-driven simulations must be distinguished from the following situations. Van Groenendaal and Kleijnen (1996, figure 1), for example, make a scatter plot of predictions versus realizations. The two coordinates of a point use the same deterministic input; different points, however, correspond with different inputs. Hence points have different expectations and variances! So it is nonsense to test the hypothesis of equal means and variances respectively (see equation 1). It seems reasonable to compute the coefficient of determination R^2 (not ρ^2 or β_0 and β_1; see equation 2), to quantify the percentage of variation 'explained' by the model. There is no statistical test statistic for R^2; it is a mathematical (not a statistical) measure; see Kleijnen (1987, p. 193). Also see Mitchell (1996).

1.2 Literature on Validation

General discussions on validation of simulation models in MS/OR can be found in all textbooks on simulation, for example, Banks and Carson (1984), Law and Kelton (1991, pp. 298-324), and Pegden, Shannon, and Sadowski (1990, pp. 133-162). A well-known article on validation is Sargent (1991). A new monograph is Knepell and Arangno (1993). Recent survey articles are Balci (1994), including 102 references, and Kleijnen (1995), including 61 references. There are also many publications outside MS/OR, for example, in agriculture (see Mitchell 1996, and Muchow and Bellamy 1991) and in the earth sciences including hydrology, geochemistry, meteorology, and oceanography (Oreskes, Shrader-Frechette, and Belitz 1994). These contemporary publications all agree that it is essential to further develop the theory on validation, because of its great importance in the practice of MS/OR.

Unfortunately, the literature gives neither a standard theory on validation, nor a standard 'box of tools'. The literature does give a plethora of philosophical theories, statistical techniques, and software practices. The emphasis of the present article is on statistical techniques.

It might be argued that statistical techniques are not appropriate in validation. Statistical techniques, however, have the advantage of yielding reproducible, objective, quantitative data about the quality of a given simulation model.

Some authors (for example, Law and Kelton 1991, p. 319) claim that, when using statistical techniques, *hypothesis tests* are inappropriate; instead they advocate confidence intervals. Hypothesis tests, however, are closely related to confidence interval procedures; see Conover (1980), Kleijnen (1974), and also Law and Kelton (1991, p. 320). Moreover, the null-hypothesis on the means of (say) X and Y may be formulated as H_0: $E(Y) = E(X) + \delta$ where δ is not necessarily zero (δ depends on the purpose of the model; also see Mitchell 1996). This paper, however, concentrates on tests with $\delta = 0$. Such tests may easily overlook 'small' differences between the real and the simulated systems (a most powerful test is the t test for either independent or dependent X and Y, provided the assumption of normality holds; distribution-free tests may be surprisingly powerful; see again Conover 1980 and Kleijnen 1974.) However, some lack of power is acceptable, if in practice only 'large' differences are important. Anyhow, testing is only part of the whole validation process (again see the references above).

Unfortunately, experience shows that the correct use of mathematical statistics in MS/OR is less simple than might be expected. It is easy to apply the wrong statistical techniques: there is much statistical software, but that software does not warn against

abuse, such as violations of statistical assumptions. On hindsight the correct use of statistics may seem easy. Indeed, Balci (1995) states: 'False beliefs exist about testing ... testing is easy ... no training or prior experience is required' (also see Mitchell 1996). This paper will provide a case in point: the wrong regression test has been used in many simulation publications (see the references in the next subsection, §1.3).

1.3 Contribution by This Paper

This paper is meant to contribute to the practice and the theory of validation (but it gives no panacea). It discusses how to validate trace-driven simulation models, emphasizing the familiar statistical technique of regression analysis, but advocating a novel test (regressing differences of simulated and real responses on sums).

Validation was interpreted above as *comparing* real and simulated outputs. More specifically, the analysts may compare the total costs (stock-carrying plus ordering plus lost-sales costs), averaged over all real and simulated days respectively.

Many years ago, Aigner (1972) already pointed out that it is wrong to expect unit slope and zero intercept, when regressing the simulated on the real outputs. He, however, focussed on econometric simulation models; he did not give the statistical test we shall propose in this paper. Years after Aigner, Harrison (1990) rediscovered that many authors still propose this bad intuitive idea. Harrison, however, discussed farming systems and synthetic models (including autocorrelations), not trace-driven discrete-event simulations; he does not propose the test we shall develop in this paper. Mayer, Stuart, and Swain (1994) challenge Harrison (1990), concluding that the old test is valid for their type of models (with autocorrelations).

Note: Aigner (1972) states that the intuitive idea dates back to Cohen and Cyert (1961). Harrison (1990, p. 184) refers to some more publications that apply this idea. Lysyk (1989) also uses this concept. Recently, the same idea was proposed in Kleijnen (1995, p. 155). So it seems high time to get rid of this concept, and to propose a better analysis. This is exactly the topic of this paper!

The essential assumption of the new test is *normally and independently distributed (n.i.d.)* outputs of the real system and the simulated system respectively. In practical simulations, however, output data may be non-stationary and autocorrelated. Unfortunately, most practitioners are familiar only with elementary statistical procedures that assume identically and independently distributed (i.i.d.) variables. Fortunately, it might be possible to derive i.i.d variables in simulation, so that it is correct to apply elementary statistical theory; for example, Law and Kelton (1991) give many examples of i.i.d. inputs and outputs, in their discussion of validation. Anyhow, in practice, simulationists often use the n.i.d. assumption, as is illustrated by the many applications (of the old test) referenced above. In general, *terminating* simulations (see Kleijnen and Van Groenendaal 1992, pp. 187-190) may give i.i.d. outputs, as is illustrated by the following queueing example.

The real and simulated systems should be observed under similar circumstances (see above); hence waiting times on a busy day in the real system should not be compared with waiting times of a simulated slow day. Those busy days may occur on (say) Saturdays. Suppose the simulation study concentrates on these days, because complaints are then most outspoken. Then there is still variation: some busy Saturdays are busier than others are. Obviously the busiest Saturday (of all Saturdays in the sample) should be compared with the busiest simulated day. These Saturdays may be assumed i.i.d.

Normal distributions may be explained by limit theorems. For example, trace data are summarized by one or a few statistics such as the average and selected quantiles. The queueing examples in Kleijnen et al. (1996) demonstrate that low traffic loads lead to normal output distributions for the average throughput time per day in terminating simulations. High traffic loads, however, give non-normality. This non-normality can be removed through transformations such as the Box-Cox transformation (which includes the logarithmic one); see Hoyle (1973). In practice the analysts can indeed test for non-normality: they can generate a large sample of simulated days.

This paper gives an academic *example* of the new and old tests, namely the validation of inventory simulation models. These models are derived from data provided by a *Monte Carlo laboratory* that uses academic simulation models (details will follow in §3). Surprisingly, the creation and use of such a laboratory seems novel in the research on *validation*.

Based on these experiments, this paper will give the following *conclusions*. (i) The old test (falsely) rejects a valid simulation model substantially more often than the nominal alpha level, whereas the new test has the correct type I error probability. (ii) The power of the new test increases, as the simulation model deviates more from the real system. (iii) Whereas in the queueing simulations in Kleijnen et al. 1996) the output should be transformed logarithmically, in the inventory system the original and the

transformed outputs give statistical performances that are very close. This holds for both the novel and the old tests.

1.4 Organization of This Paper

§2 discusses the regression analysis of simulated and real outputs in trace-driven simulations. It proves that the old test is wrong. As an alternative this section proposes to test the hypothesis that means and variances of real and simulated outputs are equal. That hypothesis is tested through a novel regression test. §3 discusses a laboratory for studying various validation tests; this laboratory uses inventory simulation. §4 discusses future research. §5 gives conclusions.

2 REGRESSION ANALYSIS

This section summarizes Kleijnen et al. (1996). Let Y_i and X_i denote simulated and real outputs respectively in observation i, with $i = 1, ..., n$; capital letters denote random variables. *Trace-driven* simulation means that X_i and Y_i are dependent; it is realistic to suppose that the linear correlation coefficient is *positive*: $0 < \rho_{xy} \leq 1$. Assume the n pairs (X_i, Y_i) are i.i.d. Finally, assume these pairs have a bivariate *normal* distribution. Denote the means by μ_x and μ_y, and the variances by σ_x^2 and σ_y^2.

We propose the following *stringent validation* requirement (assuming positive correlation between real and simulated responses): a simulation model is valid if the real and the simulated systems have *identical means* (say) μ, and *identical variances* (say) σ^2:

$$H_0: \mu_x = \mu_y = \mu \text{ and } \sigma_x^2 = \sigma_y^2 = \sigma^2 . \quad (1)$$

Because of the well-known relationships

$$\beta_1 = \rho_{xy}\sigma_y/\sigma_x; \ \beta_0 = \mu_y - \beta_1\mu_x \quad (2)$$

Equation (1) is equivalent to $\beta_1 = \rho_{xy}$ (> 0) and $\beta_0 = \mu(1 - \rho_{xy}) < \mu$.

An *ideal*, *utopian* simulation model has $X_i = Y_i$ or *perfect fit*: $\rho_{xy} = 1$. Hence, the *old* test hypothesizes that fitting the regression model $y = \beta_0 + \beta_1 x$ gives $\beta_0 = 0$ and $\beta_1 = 1$. However, if and only if $\rho_{xy} = 1$, Equations (1) and (2) together give $\beta_0 = 0$ and $\beta_1 = 1$. So in practice *the old test is erroneous*; the empirical data in §3 show that this error is serious indeed.

Note: In the previous century Galton discovered that $\rho_{xy} < 1$ causes what he called 'regression to mediocrity' in his study on parents' and children's heights; see Larsen and Marx (1986, p. 447).

Suppose the real and the simulated means are positive; in practice this condition holds for inventory costs and waiting times. This gives $0 < \beta_1 < 1$ and $0 < \beta_0 < \mu$.

Note: An application of the old test is provided by Lysyk (1989). He indeed finds an estimated slope significantly smaller than unity, and a significantly positive intercept. Since he expects a unit slope and a zero intercept, he tries to explain this phenomenon away. Another recent example is Kozempel, Tomasula, and Craig (1995, p. 231).

The *novel* test of the joint hypothesis in Equation (1) accounts for *dependence* between X and Y, as follows. Compute the n i.i.d. differences (say) $D_i = X_i - Y_i$ and sums $Q_i = X_i + Y_i$. Regress D on Q:

$$E(D \mid Q = q) = \gamma_0 + \gamma_1 q. \quad (3)$$

It is easy to prove that a common variance of the correlated normal variables X and Y implies zero correlation between their differences and sums, D and Q (this result is due to Pitman and Morgan back in 1939; the standard F test for equality of two variances does not apply; see Kleijnen 1987, p. 99). This zero correlation implies that in Equation (3) $\gamma_1 = 0$; common means of X and Y imply $E(D) = 0$ or in Equation (3) $\gamma_0 = 0$ (see equation 2). Hence the hypothesis in Equation (1) gives

$$H_0: \gamma_0 = 0 \text{ and } \gamma_1 = 0. \quad (4)$$

The analysts should *simultaneously* test the joint hypothesis in Equation (4), with an experimentwise error rate α_E not exceeding the prespecified value α. This joint test can use an F statistic; for the general formula see any textbook on regression analysis or standard regression software; for the specific formula see Kleijnen et al. (1996). (A conservative alternative to this F test is provided by the t test for $\gamma_0 = 0$ and $\gamma_1 = 0$ respectively, combined with Bonferroni's inequality.) There is an analogous F statistic for the hypothesis $\beta_0 = 0$ and $\beta_1 = 1$ in the old test.

Note: The joint hypothesis in Equation (1) may be rejected because the first sub-hypothesis ($\mu_x = \mu_y$) or the second sub-hypothesis ($\sigma_x^2 = \sigma_y^2$) is rejected. Hence, a less stringent validation requirement is that the real and simulated means are equal, but their variances may differ (the variances are then treated as nuisance parameters; the Taguchi approach, however, does consider the variance to be an important performance measure). The hypothesis of equal means can be tested by the well-known paired t test or a distribution-free test; see Conover (1980), Kleijnen (1987) and Mayer et al. (1994). This variance heterogeneity may give a slope b_1 that is lower or higher than one,

even if $0 < \rho < 1$; see Equation (2). Yet $0 < \beta_1 < 1$ still holds if (but not if and only if) $\sigma_x > \sigma_y$ (this condition means that the simulation reduces the variability, possibly because it does not account for idiosyncrasies in the real system). Common means μ implies for the intercept $\beta_0 = \mu - \beta_1\mu = \mu(1 - \beta_1)$ (see equation 2). So a simulation model with $\mu_x = \mu_y$ and $\sigma_x \geq \sigma_y$ gives simulated responses that -when regressed on real responses- result in a slope less than unity and in a positive intercept (smaller than the average simulation response).

3 AN INVENTORY LABORATORY

To illustrate the validation issues discussed in the preceding sections, it might seem illuminating to apply the old and the new tests to (say) an inventory system in practice. Suppose historical data on inputs (demands or lead times) were collected, and used to drive the simulation model, followed by the statistical tests of the preceding section. Suppose further that the simulation model were not rejected. What lesson would have been learned from such a case study? Maybe this result would only mean that the tests have not enough power. In other words, more might be learned initially from applying the statistical validation tests to a number of examples with *known* properties, so that it is possible to conclude whether rejecting a simulation model is correct or not! So instead of studying a real system, we construct a Monte Carlo laboratory that represents the following single-item inventory systems.

Demand per day is n.i.d. with mean 500 and standard deviation 50. Initially, lead time is a shifted Poisson with mean 5, that is, to a constant lead time of one we add a Poisson variable with mean four (this mean, however, will be changed below). There are lost sales (no backorders). If physical stock plus receivable orders drop below the reorder point (say) ROP, then an order is placed. ROP is selected such that a prespecified service level is realized. The order size is selected such that the total inventory costs are minimized; the formulas for ROP and EOQ (economic order quantity) are rather complicated; the exact specification, however, is unimportant for this study so we refer to Kleijnen and Van Groenendaal (1992, pp. 95-96). ROP and EOQ are fixed by the cost parameters (stock- carrying costs \$1 per day per unit, lost-sales costs \$100 per unit, and ordering costs \$10,000 per order).

Trace-driven simulation means that the simulation and the real system share some input; we decided to use the same *demand* history. If we used a trace with data on both demand and lead times, then the simulation model would be *perfect* (see §2). So we suppose that the analysts do not know the historical lead times; instead they use a distribution function for these times.

We assume *terminating simulation*: a simulation run ends after 365 days have been simulated. Each run starts with the same initial stock; no orders are on their way. (On hindsight, it would have been more realistic to take the stock at the end of one year as the initial stock for next year; but this example is only a laboratory anyhow.)

The real and simulated outputs are the real and simulated total inventory costs per year (denoted by X and Y). We suppose that the analysts have $n = 10$ years of data available (we also experimented with 25 years of data, but this situation gave the same qualitative conclusions). A higher sample size n increases the power of the validation tests. We use three classic values for the *type I error rate* of the validation tests: α is 0.01, 0.05, and 0.10. Obviously, a higher α increases the power of the validation tests. To reduce information overload and to save space, we shall present results for a single α value, namely 0.10.

Note: To decrease both the type I and the type II error probabilities, the analysts might increase the sample size n (also see Balci 1994). In practice, however, the number of observations on the real system is usually fixed (and small).

We take 500 macro-replications to estimate the performance of the tests; by definition, each macro-replication either rejects or accepts a specific simulation model (resulting in a binomial variable).

Some of our experiments give estimates of the *type I* error of the validation tests: in these experiments we have the analysts use the correct lead time distribution with the correct parameter (so the mean is 5). Yet their simulation is not perfect, since they use *pseudorandom number streams* that differ from the steams used by the 'real' system when sampling lead times (not demands). Both streams use Turbo Pascal's standard generator (multiplier 134775813, additive constant 1, multiplier 2^{32}). Hence the real and simulated output variables X and Y have the same distribution so the hypothesis in Equation (1) holds. Yet the realized outputs x and y differ, so the realized correlation r_{xy} is smaller than one. The probability of a validation test rejecting this (valid) simulation model should be α.

Our computer programming ensures that lead times and demands use *non-overlapping* pseudorandom number streams (the demand history, once sampled, is saved). We select the seed through the computer clock.

To study the *type II* error of the validation tests,

we create a *gap* between on one hand the 'reality' of the laboratory and on the other hand the simulation model of the analysts. This gap implies that the means and/or variances of the real and simulated outputs X and Y are different. There are infinitely many ways to create such gaps; we select a few ways, as follows.

We have the analysts use the *wrong mean lead time*. Rather arbitrarily, we have them use a mean lead time ranging between three and seven. We increase means with *steps* of size 0.2.

Both the old and the new validation tests operate on the same data (X_i, Y_i); this improves the comparison of the two tests.

This design turns out to make the laboratory give clear results: see Figure 1. This figure shows that the old test (falsely) rejects a valid simulation model substantially more often than the nominal α value, whereas the new test has the correct type I error probability.

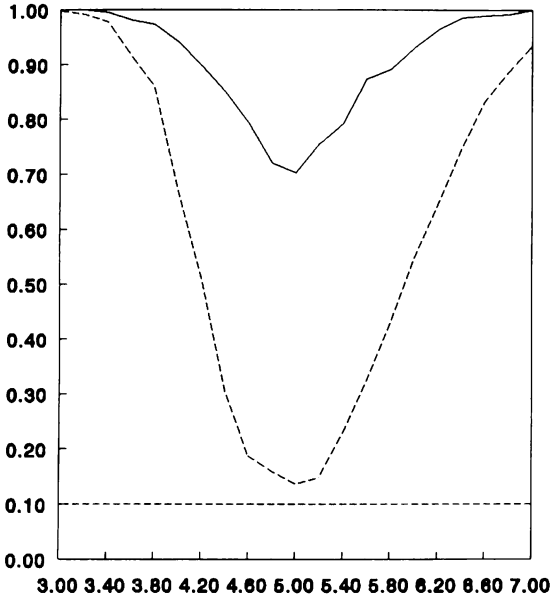

Figure 1: Estimated Probability of Rejecting the Simulated System, as a Function of the Deviation between the Means of Simulated and Real Lead Times, for the New and Old Validation Tests with $\alpha = 0.10$

Note: The power of any statistical test can be maximized over the whole domain of the parameters being tested, by simply always rejecting the null-hypothesis (that is, the simulation model is always rejected). Obviously, such a procedure is inferior. Therefore the first condition for any test is that its type I error probability is acceptable.

The figure further shows that as the analysts' error increases, the power of the test increases too. This (good) behavior is found for all type I error rates α and sample sizes n studied (but not displayed in the figure).

The figure is *asymmetric*: a mean simulated lead time (say) one unit too high has a very different effect on the total inventory costs than has a mean lead time one unit too low (lost sales against stock carrying). We also studied mean simulated lead times that are wrong on a logarithmic scale; they gave a more symmetric figure (not shown).

Notice that the analysts make *no specification errors*, when specifying the distribution type of the demands. Kleijnen et al. (1996) do study specification errors in their queueing simulations.

The analysts might apply a *normalizing transformation* to the outputs, such as the Box-Cox transformation. Such a transformation may make the new test better realize the prespecified type I error probability. In this inventory laboratory, however, the logarithmic transformation hardly affects the estimated performance of the two tests: the output is a sum, so apparently some limit theorem applies.

Note: The old validation test showed a certain 'perverse' behavior in the queueing simulations reported in Kleijnen et al. (1996); also see Harrison (1990, p. 187). In the inventory simulations, however, this behavior is not found.

4 FUTURE RESEARCH

Topics that require more research are:

(i) The novel test assumes *n.i.d.* observations on the real and simulated outputs (and so does the old test). How can this assumption be satisfied in simulations with *autocorrelations* and *time trends*? Autocorrelations might be removed through batching and similar approaches, which are popular in simulation (see Kleijnen and Van Groenendaal 1992, and also Mayer et al. 1994, pp. 99-100). Time trends might be removed through techniques used in econometrics; also see Barlas (1989, p. 68), who 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'. Also see the use of 'differencing' in Box and Jenkins (1976, pp. 378-379). In other words, the academic examples in this paper and its companion paper (Kleijnen et al. 1996) need to be supplemented with practical applications.

(ii) A specific type of non-normality, namely *binary output variables* may be important in practice. An example is the probability of buffer overflow.

(iii) The proposed statistical test of trace-driven simu-

lations is only part of the total validation and verification (V & V) process. This test needs to be incorporated in this total process.

5. CONCLUSIONS

This article focussed on *statistical hypothesis tests* for the validation of *trace-driven* simulations. It proved that it is *wrong* to expect unit slope and zero intercept when regressing simulated on real outputs. Therefore this paper applied a *novel* test: regress the differences of simulated and real responses on their sums.

Both tests were evaluated and illustrated by applying them in a *Monte Carlo laboratory* with academic inventory systems.

These experiments gave the following *conclusions*. The *old* test rejects a *valid* simulation model substantially more often than the α value indicates. The novel test does not reject a valid simulation model too often (its type I error probability equals the nominal value α). The power of the new test increases, as the simulation model deviates more from the real system.

REFERENCES

Aigner, D.J. (1972), 'A note on verification of computer simulation models', *Management Science*, 18, 11, 615-619.

Balci, O. (1995), 'Principles of simulation model validation, verification, and testing', *International Journal in Computer Simulation*.

Balci, O. (1994), 'Validation, verification, and testing techniques throughout the life cycle of a simulation study', in: *Annals of Operations Research*, 1994.

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

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

Box, G.E.P. and Jenkins, G.M. (1976), *Time series analysis; revised edition*, Holden-Day, San Francisco.

Cohen, K.J. and Cyert, R.M. (1961), 'Computer models in dynamic economics', *The Quarterly Journal of Economics*, 75, 112-127.

Conover (1980)

Harrison, S.R. (1990), 'Regression of a model on real-system output: an invalid test of model validity', *Agricultural Systems*, 34, 183-190.

Hoyle, M.H. (1973), 'Transformations; an introduction and a bibliography. *International Statistical Review*, 14, 2, 203-223.

Kleijnen, J.P.C. (1995), 'Verification and validation of simulation models', *European Journal of Operational Research*, 82, 1, 145-162.

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

Kleijnen (1974), *Statistical techniques in simulation, volume I*, Marcel Dekker Inc., New York.

Kleijnen, J.P.C., Bettonvil, B., and Van Groenendaal W. (1996), Validation of simulation models: a novel regression test. *Management Science* (accepted).

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

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

Kozempel, M.F., Tomasula, P. and Craig, J.C. (1995), 'The development of the ERRC food process simulator', *Simulation; Practice and Theory*, 2, 4-5, 221-236.

Larsen, R.J. and Marx M.L. (1986), *An Introduction to Mathematical Statistics and its Applications*. Prentice-Hall, Englewood Cliffs, New Jersey.

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

Lysyk, T.J. (1989), 'Stochastic model of Eastern spruce budworm (lepidoptera: tortricidae) phenology on white spruce and balsam fir', *Journal of Economic Entomology*, 82, 4, 1161-1168.

Mayer, D.G., Stuart, M.A., and Swain, A.J. (1994), Regression of real-world data on model output: an appropriate overall test of validity. *Agricultural Systems*, 45, 93-104.

Mitchell, P.L. (1996), Misuse of regression for empirical validation of models. *Agricultural Systems* (accepted).

Muchow, R.C. and Bellamy, J.A., editors (1991), *Climatic Risk in Crop Production*, CAB International, Wallingford.

Oreskes, N., Shrader-Frechette, K. and Belitz, K. (1994), Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, 263, 641-646.

Parks, S.K. and Miller K.W. (1988), Random number generators: good ones are hard to find. *Communications of the ACM*, 31, 10, 1192-1201.

Pegden C.P., Shannon R.E. and Sadowski R.P. (1990), *Introduction to Simulation using SIMAN*,

McGraw-Hill, New York.

Sargent, R.G. (1991), Simulation model verification and validation, *Proceedings of the 1991 Winter Simulation Conference*, 37-47.

Van Groenendaal, W. and Kleijnen J.P.C. (1996) Regression metamodels and design of experiments, *Proceedings of the 1996 Winter Simulation Conference* (edited by J.M. Charnes, D.J. Morrice, D.T. Brunner, and J.J. Swain).

AUTHOR BIOGRAPHY

JACK P.C. KLEIJNEN is Professor of Simulation and Information Systems in the Department of Information Systems and Auditing; he is also associated with the Center for Economic Research (CentER). Both the Department and the Center are within the School of Management and Economics of Tilburg University (KUB) in Tilburg, Netherlands. He received his Ph.D. in Management Science at Tilburg University. His research interests are in simulation, mathematical statistics, information systems, and logistics. He published six books and more than 130 articles; lectured at numerous conferences throughout Europe, the USA, Turkey, and Israel; he was a consultant for various organizations; and is a member of several editorial boards. He spent some years in the USA, at different universities and companies. He was awarded a number of fellowships, both nationally and internationally.

BERT BETTONVIL is Assistant Professor of Simulation in the Department of Information Systems and Auditing of the School of Management and Economics of Tilburg University (KUB). He received his Ph.D. at this university; the topic was sequential bifurcation. His main research interest is simulation, especially its programming and statistical aspects.

WILLEM J.H. VAN GROENENDAAL is Assistant Professor in the Department of Information Systems and Auditing, and Fellow at the Center for Economic Research (CentER). Both the Department and the Center are within the School of Management and Economics of Tilburg University in Tilburg, Netherlands. He is also a consultant for several international development agencies. He has a Ph.D. from Tilburg University. His research interests are in simulation, decision support, and investment analysis.