Tilburg University

Validation of Simulation Models

Kleijnen, J.P.C.; Bettonvil, B.W.M.; van Groenendaal, W.J.H.

Publication date:
1996

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):

Kleijnen, J. P. C., Bettonvil, B. W. M., & van Groenendaal, W. J. H. (1996). *Validation of Simulation Models: Regression Analysis Revisited*. (CentER Discussion Paper; Vol. 1996-07). Operations research.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

**VALIDATION OF SIMULATION MODELS:
REGRESSION ANALYSIS REVISITED**

**JACK P.C. KLEIJNEN, BERT BETTONVIL, and
WILLEM VAN GROENENDAAL**

Department of Information Systems and Auditing (BIKA)/Center for Economic Research (CentER), Tilburg University (Katholieke Universiteit Brabant), 5000 LE Tilburg, Netherlands; e-mail: kleijnen@kub.nl; fax: +3113-663377

This paper proves that it is wrong to require that regressing a model's outputs on the observed real outcomes give a 45° line through the origin (unit slope, zero intercept). Therefore this paper proposes an alternative requirement: the responses of the model and the real system should have the same means and the same variances. To test whether this requirement is satisfied, a novel statistical procedure is derived. This procedure regresses the differences of simulated and real responses on their sums. The old and the new procedures are investigated in an extensive Monte Carlo experiment that simulates queueing systems. The conclusions of this experiment are that (i) the old test rejects a valid simulation model substantially more often than the novel test does; (ii) the intuitive test shows 'perverse' behavior in a certain domain: the worse the simulation model, the higher its probability of acceptance; and (iii) the novel test does not reject a valid simulation model too often (its type I error probability is correct), provided the queueing response is transformed logarithmically.

(Keywords: TRACE-DRIVEN, CORRELATION, PAIRED OBSERVATIONS, GOODNESS-OF-FIT, SIMULTANEOUS TESTS, POWER)

1. Introduction

This introductory 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?

Sub (i): definition of validation

This paper's definition of validation follows the classic simulation 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 the validation issues discussed in this paper, consider the following *practical problem*. The management of (say) a factory is confronted with complaints about excessively long throughput times. One solution that comes to mind is increasing the number of machines, called 'servers' in queueing theory. To evaluate this solution, management wants to know *how much* the throughput times will decrease when they add one server to the existing single server. To answer their question, the Management Science/Operations Research (MS/OR) specialists build a simulation model that represents different numbers of servers. Before using this model to advise management, the MS/OR experts should validate their model; that is, they should determine whether the model is an accurate representation of the factory's queueing system. Obviously, validation should not

aim at a *perfect* model: the perfect model would be the factory (the real system) itself. So, validation may be interpreted as comparing *data* on the real and the simulated systems. Those data pertain to the inputs (such as customer or job arrival times and service duration times) and outputs (for example, job throughput times and machine idle times). Comparing the output data of the real and simulated systems makes more sense if both systems are observed under *similar circumstances or scenarios*: the throughput times on a busy day in the real factory should not be compared with the throughput times of a simulated slow day. Those busy days may occur on (say) Mondays. Suppose the simulation study concentrates on these days, because complaints are then most outspoken. Then there is still variation: some busy Mondays are busier than others are. Obviously the busiest Monday (of all Mondays in the sample) should be compared with the busiest simulated day (also see §2.1).

Hence, for validation purposes the analysts should feed real-world input data into the model, in *historical* order. This is called *trace driven* simulation in computer performance modeling; we shall use this term throughout this paper. After running the simulation program, the analysts obtain a time series of simulation output; they compare that time series with the historical time series for the output of the existing system.

Sub (ii): literature on validation

General discussions on validation of simulation models 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. 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.

Statistical techniques have the advantage of yielding reproducible, objective, quantitative data about the quality of a given simulation model. 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, in another article Balci (1995) states: 'False beliefs exist about testing ... testing is easy ... no training or prior experience is required'. This paper will provide a case in point.

Sub (iii): contribution by this paper

This paper is meant to contribute to the practice and the theory of validation. It discusses in detail how to validate simulation models (but also other models, such as econometric models), emphasizing the familiar statistical technique of regression analysis, but introducing a novel test (that regresses the differences of simulated and real responses on their sums; see later).

Validation was interpreted above as *comparing* real and simulated output time series. More specifically, the analysts may compare the *average* throughput times, computed over all real and simulated Mondays 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 tests 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, not trace-driven queueing simulations, and does not propose the test we shall develop in this paper. 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) is one more researcher who uses this intuitive idea. Recently, the same idea was proposed in Kleijnen (1995, p. 155). So it seems high time to get rid of this idea, and to propose a better analysis. This is exactly the topic of this paper!

This paper derives a better statistical procedure. The paper also applies this procedure to the validation of queueing simulation models. These models are derived from data provided by a *laboratory*, namely a Monte Carlo experiment that uses M/M/1 and M/G/1 simulation models. The creation and use of such a laboratory seems novel in the research on validation.

This paper gives the following *conclusions*. (i) The probability of rejecting a correct simulation model is substantially higher with the intuitive test than with the novel test. (ii) The intuitive test shows *perverse* behavior in certain parts of the domain: the worse the simulation model is, the higher its probability of acceptance! (iii) The novel test has a type I error probability that equals the nominal value α , provided the queueing response is transformed logarithmically. (iv) In queueing studies the simulation model might use service times with a coefficient of variation that hides a misspecified service distribution.

Sub (iv): organization of this paper

§2 discusses how to regress simulated outputs on real outputs in trace-driven queueing simulations. It is proven that the intuitive idea is wrong indeed. As an alternative this paper proposes to test for equal means and equal variances of real and simulated outputs. The power of the various tests is also discussed briefly. §3 discusses a laboratory for studying various validation tests; this laboratory uses queueing simulation. §4 gives a summary and conclusions.

2. Regression Analysis

2.1 Introduction

When validating a simulation model, it is a problem that simulation output data may form a non-stationary time series, whereas most practitioners are familiar with elementary statistical procedures that assume *identically and independently distributed* (i.i.d.) variables. Fortunately, it is easy to derive i.i.d variables in simulation, so that elementary statistical theory can be applied; this is demonstrated by the factory example, as follows.

The simulated factory may give abundant data: each individual job has a throughput time, and there may be many jobs on each day; many days can be simulated with real input data (trace driven simulation). It is practical to summarize the data per day by a few statistics, such as the average, the median (50% quantile), and the percentage of jobs waiting longer than (say) one minute. This paper concentrates on the *average* (for the other statistics see Kleijnen 1987, pp. 28-45 and Kleijnen and Van Groenendaal 1992, pp. 195-197).

So let Y_i and X_i denote the average throughput time on the i th Monday in the

simulated and the real system respectively; random variables are denoted by capitals, whereas their values are small letters. Suppose that n days are both simulated and observed in reality, so the subscript i runs from 1 through n .

We consider a *terminating* simulation. So, the averages Y_i and X_i are not computed from a steady state time series of individual throughput times. Instead, they are calculated from the individual throughput times of all jobs arriving between (say) 8 a.m., when the factory opens, and 6 p.m., when no new jobs are accepted; all jobs are finished before the next day. So each day includes a start-up, transient phase. (Also see Kleijnen and Van Groenendaal 1992, pp. 187-190.)

Hence the simulated averages Y_i are i.i.d., and so are the 'real' averages X_i . Suppose further that the historical arrival *or* service times are used to drive the simulation model (logical *or* is meant). Statistically, trace-driven simulation means that the members of the *pair* (X_i, Y_i) are *dependent* (and hence are correlated; see next paragraph). The n pairs (X_i, Y_i) are i.i.d.

Suppose that on (say) the fourth Monday in the sample of real data, the average throughput time is relatively high; that is, that value is higher than expected on a Monday (because interarrival times were relatively short or service times were relatively long on that day): $X_4 > E(X)$. Then it seems reasonable to require that on that day the simulated average (which uses the same arrival or service times) is also relatively high: $Y_4 > E(Y)$. Statistically, this implies that X_i and Y_i should have a *positive linear correlation coefficient* (say) ρ : $0 < \rho \leq 1$.

Intuitively, an *ideal* simulation model would have all its responses equal to the real responses (even though the simulated structure is a simplification of the real structure): $X_i = Y_i$ with $i = 1, \dots, n$. Such a utopian simulation model would imply *perfect fit*: $\rho = 1$.

Moreover, the fitted regression line $y = \beta_0 + \beta_1 x$ would have unit slope and zero intercept:

$\beta_1 = 1, \beta_0 = 0$. *What is wrong with this intuition?*

2.2 Statistical Analysis of Wrong Intuitive Idea

To evaluate the intuitive idea presented in the preceding subparagraph (§2.1), it is necessary to formulate a *metamodel* of the relationship between X (real response) and Y (simulated response). As such a model we propose the *bivariate normal* distribution, denoted by N_2 :

$$(X, Y) \sim N_2(\boldsymbol{\mu}, \text{cov}(X, Y)) \quad (1)$$

where $\boldsymbol{\mu}$ denotes the vector of means $(\mu_x, \mu_y)'$, and the covariance matrix is

$$\text{cov}(X, Y) = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \quad (2)$$

Observe that this distribution involves five parameters: $\mu_x, \mu_y, \sigma_x, \sigma_y$, and ρ . Non-normality will be discussed in §3.2 (which covers results of the experiments with the laboratory).

It is well-known that the bivariate normal distribution implies a *linear* relationship between the conditional mean of one variable and the value of the other variable:

$$E(Y | X = x) = \beta_0 + \beta_1 x \quad (3)$$

where the regression parameters are given by

$$\beta_1 = \rho\sigma_y/\sigma_x, \beta_0 = \mu_y - \beta_1\mu_x; \quad (4)$$

we select the simulated variable Y as the dependent variable, because the real system determines the simulated system.

Further it is well-known that the *fitting error* (say) U is defined by

$$Y_i = E(Y_i | x) + U_i \quad (5)$$

with

$$U \sim N(0, \sigma_y^2(1 - \rho^2)). \quad (6)$$

We propose the following *stringent validation requirement*: a simulation model is valid if and only if the real and the simulated systems have *identical means and variances* respectively, and the real and the simulated responses are *positively* correlated ($0 < \rho < 1$):

$$\mu_x = \mu_y = \mu \wedge \sigma_x^2 = \sigma_y^2 = \sigma^2 \wedge \rho > 0. \quad (7)$$

This requirement is equivalent to the following requirement for the regression parameters (see eq. 4):

$$\beta_1 = \rho > 0 \wedge \beta_0 = \mu(1 - \rho) < \mu. \quad (8)$$

The case $\rho = 1$ was called a *utopian* situation (see §2.1). But only in that unrealistic case does (8) give

$$\beta_0 = 0 \wedge \beta_1 = 1, \quad (9)$$

which was the intuitive idea! (If $\rho = 1$ then the covariance matrix in eq. 2 becomes singular: degenerated case.)

Suppose the real and hence the simulated means are positive; this condition certainly holds for queueing systems such as the factory example. Then (7) and (8) give

$$0 < \beta_1 < 1 \wedge 0 < \beta_0 < \mu. \quad (10)$$

Comparing (9) and (10) proves that *the intuitive idea is wrong*. We shall also give

empirical data that illustrate how (9) gives wrong tests results (see §3, which covers the queueing laboratory).

2.3 Correct Statistical Tests

It is well-known that if the pair (X, Y) is *normally* distributed, then it is optimal to use Ordinary Least Squares (OLS) to estimate the intercept and slope in the regression model. OLS gives the estimates (say) b_0 and b_1 , which are defined strictly analogous to the population parameters β_0 and β_1 in (4): $b_0 = \bar{y} - b_1\bar{x}$, etc. The sample correlation coefficient R is defined strictly analogous to ρ ; it measures the resulting goodness of fit.

We propose the following validation procedure.

Step 0: The analysts might perform a *preliminary* test concerning the weak null-hypothesis (also see eq. 4), which is a subhypotheses within the composite hypothesis in (7):

$$H_0: \rho \leq 0 \Leftrightarrow \beta_1 \leq 0. \quad (11)$$

It is well-known that to test for zero correlation, the sample correlation coefficient R may be transformed:

$$t_{n-2} = r(n-2)^{1/2}/(1-r^2)^{1/2} \quad (12)$$

where t_{n-2} is the standard symbol for Student's statistic with $n-2$ degrees of freedom: there are n observations (namely, n pairs), and two estimated regression parameters (b_0 and b_1). Computing this test is easy; there is also ample regression software to do that calculation.

In practice, this preliminary step is probably skipped; the analysts proceed

immediately to the following step. So we shall not further examine the hypothesis in (11); however, we shall use the test statistic that was formulated in (12) (see the text between eqs. 13 and 14).

Step 1: Formulate a *stringent* hypothesis: the real and the simulated systems have *identical means and variances*; see the remaining two subhypotheses in (7). How to test this composite hypothesis, while accounting for the dependence between X and Y ?

Testing for identical *means* of the dependent X and Y is easy: simply compute their Difference (say) D . The n differences are i.i.d. (because the pairs are i.i.d.). If X and Y are (approximately) bivariate normal, then the linear transformation D is (even more) normally distributed. Hence the first part of the composite hypothesis, can be tested by the *paired t statistic*:

$$t_{n-1} = (\bar{d} - \mu_d)/(s_d/\sqrt{n}) \quad (13)$$

where $\bar{d}$ denotes the sample average and s_d represents the classic estimated Standard deviation of D .

Testing for identical *variances* of the dependent X and Y is not a standard problem: the F test does not apply. Yet this problem was solved by Pitman and Morgan back in 1939 (see Kleijnen 1987, p. 99). Their test requires that besides the differences $D_i = X_i - Y_i$, the Sums (say) $S_i = X_i + Y_i$ are computed. It is simple to prove that the variances of X and Y are equal if and only if their difference and sum are uncorrelated. To test for zero correlation between D and S , we apply the t test of (12), replacing the symbol r by (say) $r_{d,s}$.

There is a statistical complication: the composite hypothesis consists of two subhypotheses, which are individually tested at a *type I error rate* of α . Then the resulting so-called *experimentwise error rate* (say) α_E when simultaneously testing two hypotheses,

exceeds the value α . This problem might be solved by testing the individual sub-hypotheses at $\alpha = \alpha/2$: Bonferroni's inequality implies that $\alpha_e \leq \alpha$ (see Miller 1981).

However, we propose a more *powerful, exact simultaneous test*. Regress D on S (analogously to the way Y was regressed on X ; see eq. 3):

$$E(D | S = s) = \gamma_0 + \gamma_1 s \quad (14)$$

where (analogous to eq. 4)

$$\gamma_1 = \rho_{d,s} \sigma_d / \sigma_s, \gamma_0 = \mu_d - \gamma_1 \mu_s. \quad (15)$$

A common variance of X and Y implies zero correlation between D and S (see above):

$\rho_{d,s} = 0$ or $\gamma_1 = 0$ (see eq. 15). Common expectations of X and Y implies a zero mean for their difference D : $\mu_d = 0$. Together, $\gamma_1 = 0$ and $\mu_d = 0$ imply a zero intercept: $\gamma_0 = 0$. So the hypothesis of real and simulated responses having identical means and variances (see eq. 7) gives

$$H_0: \gamma_0 = 0 \wedge \gamma_1 = 0. \quad (16)$$

Simultaneously testing the two regression parameters γ_0 and γ_1 is a standard problem. Because we need the corresponding formulas later, we now show how to solve this standard problem.

Compute the Sum of Squared Errors (say) SSE without and with the composite hypothesis in (16), which correspond with the *full* and the *reduced* regression model respectively. In other words, first compute

$$SSE_{full} = \sum_{i=1}^n (D_i - \hat{D}_i)^2 \quad (17)$$

where the full regression model for D on S in (14) with estimators C_0 and C_1 for γ_0 and γ_1 implies

$$\hat{D}_i = C_0 + C_1 S_i. \quad (18)$$

The reduced regression model incorporates the null-hypothesis in (16), which means $\hat{d}_i = 0$. So analogous to the SSE_{full} in (17) compute

$$SSE_{reduced} = \sum_{i=1}^n D_i^2. \quad (19)$$

It is well-known that the following expression is an F statistic with degrees of freedom equal to two (number of regression parameters in composite hypothesis, eq. 16), and $n - 2$ (n observations; two estimated regression parameters, C_0 and C_1):

$$F_{2, n-2} = [(n-2)/2][SSE_{reduced} - SSE_{full}]/SSE_{full}. \quad (20)$$

Computing this F test is easy; there is also ample regression software to do that calculation. (Details can be found in Kleijnen 1987, pp. 156-157).

If this F is significantly high, the analysts should reject the composite hypothesis in (16). In that case they should conclude that the simulation model does not meet the stringent validation requirement formulated in (7). What can the analysts do next?

The composite hypothesis is 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 treated as nuisance parameters). This hypothesis can be tested by the paired t statistic defined in (13).

This variance heterogeneity may give a slope b_l that is lower or higher than one, even if $0 < \rho < 1$; see (4). Yet the first part of (10) still holds ($0 < \beta_l < 1$) 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 (see eq. 4):

$$\beta_0 = \mu - \beta_1 \mu = \mu(1 - \beta_1). \quad (21)$$

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).

To compare the novel and the intuitive hypothesis tests, we need to formulate the joint test of the intuitive hypothesis (unit slope, zero intercept; see eq. 9). The test statistic is analogous to (20). So SSE_{full} now follows from $Y_i - \hat{Y}_i$ and $SSE_{reduced}$ follows from $Y_i - X_i$ (also see Kleijnen and Van Groenendaal 1992, pp. 209-210).

An example of testing the intuitive hypothesis is provided by Lysyk (1989). He finds an estimated slope significantly smaller than unity, and a significantly positive intercept; see (21). 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).

2.4 Type I and Type II Errors

It is well-known that when testing the validity of a simulation model, the analysts can make either a 'type I' or a 'type II' error. So they may reject the simulation model, while this model is valid: type I or α error. Or they may accept the simulation, while this simulation is not valid: type II or β error. The probability of a type II 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*; see Balci

(1994).

Note that 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 power of the paired t test of $H_0: \mu_d = 0$ increases, as the *model specification error* $|\mu_d|$ increases. For example, as $|\mu_d|$ goes to infinity so does $|t_{n-1}|$ in (13); hence the simulation model is rejected for any sample size n and for any type I error probability α (together, n and α fix the critical value $t_{n-1; \alpha/2}$).

The probability of a type II error increases as α decreases, given a fixed number of simulated days n : 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 selection of a particular value for α is problematic. Popular values are 0.01, 0.05, and 0.10. These values will be investigated later (see §3).

To decrease both the type I and type II error probabilities, it is necessary to increase the *sample size* n . In practice, however, the number of observations on the real system is usually fixed (and small).

Finally, consider the test of the intuitive hypothesis (unit slope, zero intercept); see (9). Harrison (1990, p. 187) points out that this test may show the following perverse behavior. When the simulation outputs (Y) deviate substantially from the real outputs (X), then the correlation coefficient ρ is small, so the fitting error U in (5) has a large variance; see (6). But then SSE_{full} goes up, so $F_{2, n-2}$ goes down. Consequently, the worse the simulation model is, the higher the probability of not rejecting it! The paired t test in (13)

does not have this bad behavior. We shall come back to this 'perverse' phenomenon in the next section.

3. Queueing Laboratory

3.1 Creating a Queueing Laboratory

To illustrate the validation issues discussed in the preceding section, it might seem illuminating to test the traditional and the novel hypotheses (see eqs. 9 and 16) for a factory in practice, as the example in §1 suggested. Suppose historical data on inputs (arrival or service times) were collected, and used to drive the simulation model, followed by the statistical procedures 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. More can be learned from applying the statistical validation procedures to a number of examples with *known* properties, so that it is possible to conclude whether rejecting a hypothesis is correct or not!

So instead of studying a real factory, we construct a *laboratory*, as follows. In practice, models are not perfect representations of reality, because certain parts of the real system can not be observed: these parts are treated as *black boxes* (also see the seminal book, Zeigler 1976). For example, suppose the analysts observe jobs enter and leave a queueing system, such as a factory. In other words, the analysts record the interArrival times (say) a_t and the departure or Exit times e_t of the jobs t with $t = 1, 2, \dots$. The analysts can not observe what is happening inside the queueing system: the Service time s_t can not

be measured. The analysts do measure the throughput times (say) z_t . Notice that we assume that the analysts are able to identify the jobs; in the laboratory, jobs are identified by their 'name' t . Suppose further that upon studying the sequence of inputs a_t and outputs e_t , the analysts conclude that the queuing system uses the first-in-first-out (FIFO) queueing priority. Suppose finally that the analysts find out that there is a single server inside the system (to find this out, they might talk to the factory's management).

The analysts know that throughput time equals the sum of service and Waiting times (say) w_t . This yields the identities

$$Z_t = E_t - A_t = W_t + S_t. \quad (22)$$

Further, the analysts know that the dynamics of a single-server system with FIFO is determined by

$$W_t = \max(0, W_{t-1} + S_{t-1} - A_t, 0) \quad (23)$$

where $W_1 = 0$, since the queueing system starts in the empty state.

From the variation in the observed interarrival times (a_t) and exit times (e_t) the analysts conclude that interarrival times and service times are stochastic variables:

$$A_t \in F_a(t), S_t \in F_s(t) \quad (24)$$

where the argument t in the distribution functions $F(t)$ indicates that the arrival and service times may form time series.

We assume that *the* response of the real system is X , the average value of the time series of T throughput times per day:

$$X = \sum_1^T Z_t/T. \quad (25)$$

The analysts get data over n days, which yields X_i with $i = 1, \dots, n$ (see §2.1).

Suppose the analysts model the 'real' system (our laboratory) as an $M/M/1$ model: interarrival and service times form Poisson processes with parameter λ_a and λ_s respectively, and there is one server (FIFO is the default priority option). Let a tilde denote variables used in the simulation model (not the laboratory) and let $Ne(\lambda)$ denote the Negative exponential distribution with rate λ . Then the analysts do not use (24) but

$$\tilde{A}_t \in Ne(\lambda_{\tilde{a}}), \tilde{S}_t \in Ne(\lambda_{\tilde{s}}). \quad (26)$$

Analogous to (25), *the* response of the simulated system is the average value of the time series of simulated throughput times, per day:

$$Y = \sum_1^T \tilde{Z}_t / T \quad (27)$$

where $\tilde{Z}$ is defined analogous to (22) as

$$\tilde{Z}_t = \tilde{E}_t - \tilde{A}_t = \tilde{W}_t + \tilde{S}_t. \quad (28)$$

Trace driven simulation means that the simulation uses the arrival times of the real laboratory system: $\tilde{a}_t = a_t$ and (25) and (27) use the same symbol T .

When modeling a simple system such as our laboratory, the analysts make no modeling errors when specifying the *structure* of the system: both the laboratory and the simulation are single-server FIFO systems. However, the analysts may make errors when specifying the *distribution type* of the service times, which they could not observe. Above we assumed that the analysts specify a Markovian (symbol: M) service process. Actually, the laboratory may have a general (symbol: G) service time distribution; that is, the laboratory may be M/G/1, whereas the simulation is M/M/1. For G we select an *Erlang* distribution with parameters k and λ such that $\mu_s = k/\lambda$ and $\sigma_s = k^{1/2}/\lambda$, which implies a coefficient of variation (say) $v_s = \sigma_s/\mu_s = k^{-1/2}$. (By definition, $k = 1$ corresponds with

M/M/1; $k = 2$ means summing two i.i.d. negative exponential distributions with parameter λ so $v_s = 2^{-1/2}$.) Pollaczek's formula shows the role of the *coefficient of variation* of the service times:

$$E(W_\infty) = [(\lambda_a/\lambda_s)(1 + v_s^2)]/[2\lambda_s(1 - \lambda_a/\lambda_s)] \quad (29)$$

where W_∞ denotes the waiting time in the steady state, and λ_a/λ_s denotes the traffic load (notice that $\lambda_s = 1/\mu_s$).

Even if the analysts specify the correct distribution type (that is, both the laboratory and the simulation are M/M/1), they make an *estimation error*: they use λ_s in (26), so they miss the correct value λ_s . In the experiments with the laboratory we can manipulate the magnitude of this estimation error, to see whether the validation tests do detect gross errors.

Hence, by creating a *gap* between the laboratory and the simulation, we study the *type II* error of the validation tests. This gap means that the means *or* variances of average throughput time per day differ between simulation and laboratory (see eq. 7). There are infinitely many ways to create such gaps; we select a few ways, as follows.

In the first experiments both the laboratory and the analysts use Poisson service times but with different rates. We have the analysts make estimation errors up to 0.10. So the mean service time of the laboratory is μ_s , whereas the analysts use a mean service time $\mu_{\bar{s}}$ such that $|\mu_s - \mu_{\bar{s}}| \leq 0.10$. We increase the estimation errors with steps of size 0.05. (Also see the x-axis in the various figures, which will be discussed later.)

In the next experiments the laboratory uses the Erlang distribution discussed above. We again take advantage of the laboratory character of our study; that is, first we assume that the analysts (who use the M/M/1 simulation model) estimate the parameter of their exponential service distribution such that $\mu_{\bar{s}} = \mu_s$; next we vary the magnitude of their

estimation error, such that $|\mu_s - \mu_{\tilde{s}}| \leq 0.10$.

To study the *type I* error of the validation tests, we run a number of experiments that have simulation models representing the corresponding real systems as accurately as is reasonable. Therefore these experiments will have both the laboratory and the simulation model be M/M/1 with equal service rates ($\lambda_s = \lambda_{\tilde{s}}$). Even in these cases the simulation models are not ideal: since the laboratory and the simulation use different pseudorandom numbers, the realizations s_t and $\tilde{s}_t$ are different, so x and y are different, and r is smaller than one.

In these 'type I' experiments the laboratory and the simulation have the same distribution for the throughput times, which implies $\mu_x = \mu_y$ and $\sigma_x^2 = \sigma_y^2$. So when testing the hypothesis corresponding with (16), the probability of the novel test rejecting the simulation model should be α . For the old test, the probability of a type I error cannot be estimated from these experiments, since the corresponding null-hypothesis ($\beta_I = 1 \wedge \beta_o = 0$; see eq. 9) never holds; obviously, it is possible to estimate the probability of rejecting a valid simulation model when using the old test.

It is realistic to have the analysts use a *pseudorandom number generator* that differs from the generator used in the laboratory. We have the laboratory use a 'good' generator that is taken from Parks and Miller (1988):

$$r_{i+1} = (7^5 r_i) \bmod (2^{31} - 1). \quad (30)$$

The analysts use Turbo Pascal's standard generator

$$r_{i+1} = (134775813 r_i + 1) \bmod 2^{32}. \quad (31)$$

Hence, the analysts sample service times $\tilde{s}_t$ that are certainly independent of the laboratory service times s_t ($t = 1, \dots, T$).

Both the laboratory and the analysts use exactly the same arrival times, since the analysts use the trace of arrival times. We further ensure that arrival times and service times use *non-overlapping* pseudorandom number streams: we generate the arrival time and service time *per* job. We select the seed (r_0) through the computer clock.

We wish to create several *laboratories* (situations, scenarios, environments). We select the following values.

(i) *Traffic load of queueing systems (noise): 0.5, 0.8, 0.9, and 1.0*

After fixing the interarrival rate at one ($\lambda_a = 1$), we select several service rates for the laboratory such that the traffic load (λ_a/λ_s) varies between 0.5 and 1.0. Low traffic rates yield many zero waiting times. Heavy traffic rates increase not only the mean but also the variance of the throughput time; and more noise in the outputs (X and Y) may affect the validation tests. Observe that a traffic load of 1.0 does not give exploding outputs, since the simulation is terminating.

(ii) *Number of jobs per day (normality): 10, 100, and 1000*

T_i (number of jobs on day i) is the same in the laboratory and the simulation, since the same arrival times are used (see eqs. 25 and 27). For simplicity's sake we make T_i deterministic; we select the values 10, 100, and 1000. We conjecture that a large number of jobs per day may make the response (average throughput time) normally distributed (the old and new validation tests assume normality).

(iii) *Number of days (power): 10 versus 25*

We choose the number of days on the trace, n , to be either 10 or 25. More observations increases the power of a test.

Sub (i), (ii), (iii): The total number of combinations is 24 ($= 4 \times 3 \times 2$); for example, laboratory #1 has a traffic load of 0.5, only 10 jobs per day, and only 10 days observed.

We study several classic values for the *type I error rate* of the two validation tests, namely α is 0.01, 0.05, and 0.10.

3.2 Results of the Experiments with the Laboratory

We wish to compare the *traditional* and the *novel* test procedures, applied to the hypotheses in (16) and (9). To make such a comparison we wish to estimate the type I and II error probabilities (see §2.4). For the intuitive test, however, the usual estimates do not give the customary power function, since the null-hypothesis corresponding with (9) never holds in our experiments ($p < 1$). Hence no point in the following three figures gives the estimated type I error rate for the intuitive test.

To estimate the various error probabilities we use 1,000 macro-replications of a specific laboratory. Each macro-replication either rejects or accepts the simulation model. In case of rejection we score (say) a zero; otherwise a one. This defines a *binomial* variable. Such a variable results for the traditional and the novel procedures; denote these two variables by A_0 and A_1 respectively. Because both tests use the same data (X_i, Y_i) , the variables A_0 and A_1 are correlated. Moreover, we use *several* values for α (namely, 0.01, 0.05, 0.10). So we get $A_0(\alpha)$ and $A_1(\alpha)$, which are all correlated.

To reduce information overload and to save space, we present pictures for only 2^3 of the 24 combinations studied (see §3.1), namely the combinations of *extreme* values. Further, the three α values give the same conclusions (of course, the higher the type I error is, the higher the estimated power is), so we present pictures for a single value, namely $\alpha = 0.10$ (this value gives the clearest picture, given the size of the figures and the reduction factor). The remaining pictures can be obtained -free of charge- from the

authors.

Figure 1 suggests the following *conclusions*.

- (i) The intuitive test always rejects the correct simulation model substantially more often than our novel test does.
- (ii) For a small number of jobs per day ($T = 10$) the 'estimated power function' of the intuitive test does not reach its *minimum* when the simulated system equals the real system ($\mu_s = \mu_s$). This estimated function is not symmetric. Actually, this test shows *perverse* behavior in certain parts of the domain. If the analysts use $\mu_s \gg \mu_s$, then $E(Y) \gg E(X)$, which tilts the regression line towards the intuitively expected regression line (zero intercept, unit slope) so the wrong simulation model is accepted!
- (iii) Even our novel test has a type I error probability that is higher than the nominal value α . We shall return to this issue, after conclusion (iv).
- (iv) As might be expected, the estimated power curves get better as the number of jobs per day (T) or the number of days (n) increases.

INSERT FIGURE 1

Sub (iii): We conjecture that the high type I error probability of the novel test is caused by *non-normality* of the pairs (X_i, Y_i) . We further conjecture that the marginal distributions of X_i and Y_i are non-normal, when the traffic load is high and there are only a few jobs per day. The latter conjecture is not inspired by Figure 1, but by the following reasoning.

The classic *Central Limit Theorem* (C.L.T.) does not apply to the individual throughput times, because they are autocorrelated, and they do not have constant means

and variances; for example, each *first* customer of a day has zero waiting time (which is part of the throughput time; see the definition of X in eq. 25). However, if the traffic load is very low, then most waiting times are zero and the throughput times are virtually equal to the service times. The service times are i.i.d, so the C.L.T. applies to their average. (If the service times are exponential distributed, then their sum is Erlang distributed.)

In case of serious non-normality the analysts may apply the *logarithmic transformation* to the simulation responses. This transformation has a mathematical requirement, namely the observations must be positive. Obviously, this requirement is met by throughput times (X and Y). This requirement, however, is not satisfied by the differences between real and simulated throughput times ($D_i = X_i - Y_i$). Statistically speaking, this transformation is applied when the distribution of the original response has a long tail at the right end. (Also see Kleijnen 1987.)

We conjecture that positive skewness does occur in queueing systems. In practice the analysts can test whether such skewness occurs: they can generate a large sample of simulated days, for one or more scenarios. We simulate 1,000 days, and test whether the resulting empirical distribution of queueing responses Y is Gaussian. For this test we use the chi-square statistic χ_v^2 . We find that the original response Y is indeed non-normal when the traffic load is high ($\mu_s = 1.0$) and there are only a few jobs per day ($T = 10$): $\chi_7^2 = 268.18$ ($v = 10 - 2 - 1$: 10 classes, 2 estimated parameters). The transformed response $\log(Y)$ has a smaller chi-square statistic: $\chi_7^2 = 12.52$, which is significant at a type I error rate of 0.10, but not at 0.05 ($\chi_{7;.10}^2 = 12.0$; $\chi_{7;.05}^2 = 14.1$). Repeating this study with low traffic ($\mu_s = 0.5$) and many jobs per day ($T = 1,000$) gives $\chi_7^2 = 8.52$ for Y , and $\chi_7^2 = 3.82$ for $\log(Y)$. So neither the original nor the transformed response is significantly non-normal; the transformation does reduce any minor non-normality.

The real question, however, is whether the transformation improves the statistical performance of the validation procedures. We study only four situations, again defined by high versus low traffic (μ_s is 1.0 vs 0.5), and many versus few jobs per day ($T = 1,000$ vs 10); that is, we keep the number of days fixed, since this factor does not affect normality of Y ($n = 10$ keeps computer time smaller than $n = 25$). This gives Figure 2, which clearly shows that this transformation makes the novel validation test realize the prespecified type I error probability (α). The old test is still inferior.

INSERT FIGURE 2

We assumed that the analysts may also make errors when specifying the *service time distribution* (see §3.1). We assumed that the laboratory uses the *Erlang* distribution with parameter $k = 2$, whereas the analysts use the exponential distribution. We saw (Figures 1 and 2) that it is necessary to apply the *logarithmic transformation* to the responses. We now study only two situations, varying the number of jobs per day ($T = 10$ vs 1,000); we study only low traffic ($\mu_s = 0.5$) and a low number of days ($n = 10$). This gives Figure 3.

Again this figure should be interpreted with care, since it does not give the usual power function, not even for our new test: no point on its curves gives the estimated type I error rate, since the analysts always use the wrong simulation model (they use the wrong type of distribution for the service time).

This figure does show that the old validation test remains inferior. This figure further shows that the false simulation model is always rejected more often than the prespecified α value. The 'estimated power function' of the novel test does not reach its

minimum when the simulated mean service time equals the real mean ($\mu_s = \mu_s = 0.5$), especially when there are many jobs per day ($T = 1000$). This phenomenon can be explained by Pollaczek's formula as follows. This formula shows the role of the *coefficient of variation*, $v_s = \sigma_s/\mu_s$; see (29). For the laboratory we select an *Erlang* distribution, which has variation coefficient $v_s = 1/\sqrt{2}$, whereas the analysts' M/M/1 simulation has $v_s = 1$. This coefficient occurs in the numerator of Pollaczek's formula, so we know that the simulation response $E(Y)$ is higher than the real response $E(X)$. But then the simulation gives a valid prediction of the real response, if the simulation uses an expected service time that is too low ($\mu_s \ll \mu_s$).

INSERT FIGURE 3

4. Summary and Conclusions

The Introduction (§1) mentioned that contemporary publications agree that it is essential to further develop the theory on validation, because of the great importance of validation in the practice of MS/OR.

However, the literature gives neither a standard theory, nor a standard 'box of tools'. The emphasis of this article is on statistical techniques. These techniques yield reproducible, objective, quantitative data about the quality of a given simulation model.

Unfortunately, experience shows that the correct use of mathematical statistics is less simple than might be expected. Indeed this paper proved that it is wrong to expect unit slope and zero intercept when regressing simulated on real outputs.

This paper introduced a novel test: regress the differences of simulated and real

responses on their sums.

The old and new statistical validation procedures were evaluated and illustrated by applying them to queuing simulation models. These models were derived from data provided by a laboratory that uses M/M/1 and M/G/1 simulation models where we let G stand for the Erlang distribution with parameter value two (sum of two i.i.d. exponentials).

These laboratory experiments gave the following *main conclusions*.

- (i) The old test rejects a valid simulation model very often, namely substantially more often than the novel test does.
- (ii) The intuitive test shows perverse behavior in a certain domain: the worse the simulation model, the higher its probability of acceptance.
- (iii) The novel test does not reject a valid simulation model too often, provided the queueing response is transformed logarithmically (then its type I error probability equals the nominal value α).
- (iv) In queueing studies the simulation model might use service times with a coefficient of variation that hides a misspecified service distribution.¹

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',

¹ Jos Verstegen, Ph.d. student at Wageningen Agricultural University, brought the article by Harrison (1990) to our attention. Dr. Richard Fleming of the Canadian Forest Service mentioned the article by Lysyk (1989). Drs. Hans Moors and Hans Blanc, both at Tilburg University, provided useful comments on draft versions of this paper.

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.
- Cohen, K.J. and Cyert, R.M. (1961), 'Computer models in dynamic economics', *The Quarterly Journal of Economics*, 75, 112-127.
- Harrison, S.R. (1990), 'Regression of a model on real-system output: an invalid test of model validity', *Agricultural Systems*, 34, 183-190.
- 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, 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.
- 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.
- Ma, B.N.W. and Mark, J.W. (1995), 'Approximation of the mean queue length of an

- M/G/c queueing system', *Operations Research*, 43, 1, 158-165.
- Miller, R.G. (1981), *Simultaneous Statistical Inference; Revised Edition*, Springer-Verlag, New York.
- 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.
- Popper, K.R. (1968), *The Logic of Scientific Discovery; Second Edition*, Harper & Row, New York.
- Sargent, R.G. (1991), Simulation model verification and validation, *Proceedings of the 1991 Winter Simulation Conference*, 37-47.
- Whitt, W. (1989), "Planning queueing simulations." *Management Science*, 35, no. 11, November, pp. 1341-1366.
- Zeigler, B. (1976), *Theory of Modelling and Simulation*, Wiley Interscience, New York.

Figure 1: Estimated Power Functions of Old (---) and New (—) Validation Tests with Fixed Real and Varying Simulated Mean Service Times for Three Factors: Real Traffic Load (μ_r), Jobs per Day (T), and Days (n), Given a Type I Error Rate (α) of .10.

Combination 1: $\mu_r = .5$, $n=10$, $T=10$

Combination 2: $\mu_r = .5$, $n=10$, $T=1000$

Combination 3: $\mu_r = .5$, $n=25$, $T=10$

Combination 4: $\mu_r = .5$, $n=25$, $T=1000$

Figure 1 (continued)

Combination 5: $\mu_s=1.0$, $n=10$, $T=10$ **Combination 6: $\mu_s=1.0$, $n=10$, $T=1000$** **Combination 7: $\mu_s=1.0$, $n=25$, $T=10$** **Combination 8: $\mu_s=1.0$, $n=25$, $T=1000$**

**Figure 2. Logarithmically Transformed Responses $\log(X)$ and $\log(Y)$,
Given # Days ($n=10$)
(see title of Figure 1 for remaining symbols)**

Combination 1: $\mu_e=.5$, $T=10$

Combination 2: $\mu_e=.5$, $T=1000$

Combination 3: $\mu_e=1.0$, $T=10$

Combination 4: $\mu_e=1.0$, $T=1000$

**Figure 3: Erlang Real and Exponential Simulated Service Times
with Logarithmic Transformation of Responses,
Given Real Traffic Load $\mu_r=0.5$ and # Days $n=10$
(see title of Figure 1 for remaining symbols)**

Combination 1: T=10

Combination 2: T=1000