Statistical aspects of simulation

Kleijnen, J.P.C.

Publication date:
1981

faculteit der economische wetenschappen

RESEARCH MEMORANDUM

TILBURG UNIVERSITY
DEPARTMENT OF ECONOMICS
Postbus 90135 - 5000 LE Tilburg
Netherlands

FEW
101

STATISTICAL ASPECTS OF SIMULATION: AN UPDATED SURVEY.

by

Jack P. C. Kleijnen

Information Systems Group
Department of Business and Economics
Katholieke Hogeschool Tilburg (Tilburg University)
P.O. Box 90153
5000 LE Tilburg
The Netherlands

Version 1

September 1981

This paper has been written at the request of an Editorial Board.
Hopefully this paper will be published in the open literature in 1982.

CONTENTS

STATISTICAL ASPECTS OF SIMULATION: AN UPDATED SURVEY

by JACK P.C. KLEIJNEN*)

ABSTRACT

Practical statistical techniques for the design and analysis of simulation experiments are presented. These techniques are relevant in both discrete and continuous, deterministic or stochastic simulation. To generalize and interpret the simulation output the analyst can use regression analysis. This analysis allows for interactions among factors. Actually the regression model provides a first-order or a second-order approximation to the complicated simulation model. To decide which system variants (parameter combinations) should be simulated, the analyst should apply experimental design theory. This theory makes the exploration of the simulated system much more efficient and more thorough. In the preliminary phase of the simulation experimentation special screening designs can be used to investigate hundreds of factors in relatively few runs.

In stochastic simulation there are additional problems. Several approaches are available for determining how to initialize a simulation run and how long to continue that run. These approaches result in a confidence interval for the estimated response. Both steady-state and transient behavior are examined. Special variance reduction techniques are briefly discussed; the use of common random numbers (identical seeds) is discussed in more detail.

1. Introduction

This article is meant as a survey for simulation practitioners. The reader should know basic statistical theory such as elementary regression analysis and t tests. The reader should not expect a discussion of the generation of random numbers and their use in sampling from distributions, as the simulation practitioner may rely on standard computer programs

*) Department of Business and Economics, Katholieke Hogeschool Tilburg (Tilburg University), 5000 LE Tilburg, The Netherlands.

for sampling from distributions. Another aspect not discussed is the use of statistical techniques for validating simulation models. The following statistical aspects, however, are discussed in detail:

(i) Strategic statistical problems.

Simulation can be applied in the study of many practical problems but unfortunately the results of a simulation experiment are valid only for the specific parameter values and mathematical relationships of the executed simulation program. If the user wishes to know the effects of changing a parameter or relationship, then the simulation program must be run again. An astronomical number of combinations of parameters could be formulated. Running all these combinations, however, would take too much computer time. This paper will recommend the use of experimental design theory to help in the efficient and systematic exploration of the great many system variants that could be simulated.

Apart from the technical impossibility of simulating all system variants, there remains the problem of how to interpret the great mass of data usually produced by computer programs. Understanding the simulated system is certainly not easy when confronted with reams of paper output or with numerous tables. This paper proposes to gain insight into the behavior of the simulation model by using a regression model that explains how the simulation output (y) reacts to changes in the simulation model's parameters and relationships ($x_1 \ldots x_k$). Hence regression analysis yields a metamodel or auxiliary model built on top of the simulation model, so-called hierarchical modeling. Obviously the validity of the regression (meta)model should be checked (just like the validity of the underlying simulation model should be checked in an earlier phase).

So experimental design and regression analysis may mitigate the ad hoc character of the simulation technique. These statistical methods are useful in determining the sensitivity of the simulation results to specific model assumptions (simulation model validation), and in finding optimal or satisficing values for the decision variables in the simulation (this includes the what-if approach). The number of references - both theoretical papers and practical applications - has been kept small; additional references are found in [11].

(ii) Tactical statistical problems.

Tactical problems are defined as problems arising when simulating a single, specific system variant, i.e., given the specification of all parameters and relationships the following (mutually related) statistical problems remain: How long should the simulation run be? Once the run is terminated, how can a confidence interval for the run's response be derived? Can special "tricks" be applied to reduce the variability of a response variable? Note that these tactical issues arise only in stochastic simulation models, whereas the strategic problems exist in both deterministic and stochastic simulation.

2. Regression metamodels.

Before discussing the details of regression metamodeling in simulation, we emphasize that this regression analysis is merely a formalization of the following common sense approach. In practice the analyst changes the value of a parameter; observes the resulting response; possibly he changes that parameter again; he plots the input/output combinations; fits a curve by band; and he concludes whether this parameters has an important effect on the output; next he repeats this process for another parameter. Regression analysis formalizes the hand-fitting by applying the least squares algorithm; regression analysis permits an extension into multiple dimensions (main effects and interactions of an arbitrary number of parameters can be analysed); to judge the importance of a parameter an objective significance test is used; the validity of the fitted curve is checked statistically. Note that the use of formal metamodels in simulation (and in other disciplines such as mathematical programming and World Dynamics) has been advocated by several authors, e.g. [14, 22].

Let the simulation model be briefly denoted by the function symbol f, with simulation response y, parameters z_j (j = 1,...k) and random number vector $\underset{\sim}{r}$:

$$y = f_1(z_1,\ldots,z_k, \underset{\sim}{r}) \qquad (2.1)$$

Some comments can be made on eq. (2.1). The intricate simulation model is indeed a mathematical function, i.e., for a given set of arguments it yields a unique value. Although the simulation program generates a

whole time series per output variable, this time series is usually summarized by a limited number of measures, e.g., the average, the maximum, the spectrum. This paper concentrates on a single statistic per simulation run (multiple statistics will be briefly discussed later; so also Appendix 1). Further note that the random vector $\underset{\sim}{r}$ is completely determined by its seed (initial value), say r_0. Metamodeling also applies to deterministic simulation in which case $\underset{\sim}{r}$ vanishes. The symbol z can represent not only a parameter like service rate but also a (discrete) quantitative variable like number of service stations, or a qualitative "variable" like queuing discipline. In experimental design terminology z is called a "factor". From a user's point of view factors in the model can be partitioned into decision and environmental factors. Decision factors are under the user's control. Environmental factors are not controllable but they do affect the output. Decision factors can be selected such that either the output is optimized, or a satisficing value of the output is obtained (related to H. Simon's theory and to the "what if" approach), or a fixed output value is realized (control approach). A valid model implies that the exact values of the environmental factors either are known or are not critical.

Returning to eq. (2.1), the simplest situation would permit a strictly mathematical foundation for metamodeling. If all factors z are continuous variables and $\underset{\sim}{r}$ vanishes, and the simulation model f_1 is a nicely behaving function, then a Taylor series expansion holds. A first-order approximation to eq. (2.1) is then

$$y = \beta_0 + \beta_1 \cdot z_1 + \ldots + \beta_k \cdot z_k \qquad (2.2)$$

In practice such simple situations do not occur. Therefore the analyst has to hypothesize a specific form of the metamodel and later on he has to test the validity of his metamodel. The analyst's hypothesis is inspired by mathematical reasoning, theoretical knowledge about the simulated system which indicates the important factors, intuition, etc.

Inspired by the first-order approximation of eq. (2.2) the analyst may postulate the following metamodel

$$y_i = \beta_0 + \beta_1.z_{i1} + \ldots + \beta_k.z_{ik} + e_i \quad (i = 1,\ldots,n) \qquad (2.3)$$

where e_i represents noise in simulation run i. A first-order metamodel like eq. (2.3), however, implies that a change in factor j has a constant effect on the expected response, independent of the other factors j' $(j' \neq j)$:

$$\frac{\partial\{E(y)\}}{\partial z_j} = \beta_j \qquad (j=1,\ldots,k) \qquad (2.4)$$

Eq. (2.4) also implies parallel response curves. The simplest metamodel allowing for interactions $\beta_{jj'}$ (non-parallel response curves) is shown in the next equation where for illustration purposes k is limited to three:

$$y_i = \beta_0 + \beta_1.z_{i1} + \beta_2.z_{i2} + \beta_3.z_{i3} + \\ + \beta_{12}.z_{i1}.z_{i2} + \beta_{13}.z_{i1}.z_{i3} + \beta_{23}.z_{i2}.z_{i3} + e_i \qquad (2.5)$$

Whatever metamodel the analyst starts out with, he has to test this model's validity. In the statistics literature the reader may find the lack-of-fit F-test. However, we would recommend the following procedure:

(i) Postulate a specific form for the regression metamodel, e.g., eq. (2.3) or (2.5).

(ii) Estimate the parameters β in this metamodel as follows. If the model contains q parameters, then the number of simulation runs should satisfy $n \geq q$. Determining <u>which</u> system variants (i.e. which combinations of z) should be simulated, is an experimental design problem, discussed later on. Let the experimental design yield a non-singular matrix of independent variables $\underset{\sim}{Z}$ with n rows and q columns.

Executing the corresponding simulation runs yields the n simulation responses y_i plus the corresponding standard errors s_i (section 4 shows how these standard errors are computed). The Ordinary Least Squares (OLS) estimators of $\underset{\sim}{\beta}$ are well-known. However, typical for simulation is that the covariance matrix of y (say $\underset{\sim}{\Omega}_y$) is usually a diagonal matrix $\underset{\sim}{D}$ (the independence of the responses is guaranteed if independent

seeds r_0 are used) with elements $\sigma_i^2 = E(s_i^2)$. Therefore OLS may be replaced by Estimated Weighted Least Squares (EWLS) where observation y_i is weighted with its estimated variance s_i^2 :

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{Z}' \cdot \hat{\underset{\sim}{D}}^{-1} \cdot \underset{\sim}{Z})^{-1} \cdot \underset{\sim}{Z}' \cdot \hat{\underset{\sim}{D}}^{-1} \underset{\sim}{Y} \qquad (2.6)$$

Analytically the variances of the estimators for $\underset{\sim}{\beta}$ can be derived only for known $\underset{\sim}{D}$ or for large sample sizes. Monte Carlo experiments reported in [13] show that if the σ_i^2 are estimated from at least five observations (e.g., five subruns; see section 4), then the asymptotic formula can still be used, i.e.,

$$\hat{\underset{\sim}{\Omega}}_{\underset{\sim}{\beta}} = (\underset{\sim}{Z}' \cdot \hat{\underset{\sim}{D}}^{-1} \cdot \underset{\sim}{Z})^{-1} \qquad (2.7)$$

The EWLS estimators give more accurate estimators of $\underset{\sim}{\beta}$, provided the σ_i^2 differ by a factor, say, ten. Then significant parameters β can be detected more frequently.

(iii) Validate the estimated regression model following the traditional scientific procedure, i.e., use the regression model to forecast the response y at a new setting of the simulation factors, say $\underset{\sim}{z}_{n+1}$:

$$\hat{y}_{n+1} = \underset{\sim}{z}'_{n+1} \cdot \hat{\underset{\sim}{\beta}} \qquad (2.8)$$

Compare the metamodel's prediction to the actual simulation response y_{n+1}. Only if the metamodel's prediction deviates significantly from the simulation model's result, the estimated regression model is rejected. An appropriate statistic is

$$(y_{n+1} - \hat{y}_{n+1}) / \{s^2_{n+1} + \underset{\sim}{z}'_{n+1} \cdot \hat{\underset{\sim}{\Omega}}_{\underset{\sim}{\beta}} \cdot \underset{\sim}{z}_{n+1}\}^{1/2} \qquad (2.9)$$

According to a recent Monte Carlo study [12] the statistic of eq. (2.9) may be approximated by the standard normal variable N(0,1). The following "trick" may be used to obtain as many validation runs as possible:

Suppose n simulation runs are available ($n > q$). Then remove one run, say run i ($i=1,\ldots,n$) and estimate $\underset{\sim}{\beta}$ from the remaining n-1 observations (assuming a non-singular matrix, say $\underset{\sim}{Z}_{(i)}$ remains).

Use the resulting estimator $\hat{\beta}_{(i)}$ to predict the simulation response of the run removed. Compare this prediction $\hat{y}_i$ tot the (removed) actual response y_i, using the statistic of eq. (2.9).Next remove a different run i' and add the previously removed run i ($i \neq i'$). Estimate β from the resulting n - 1 observations, etc. This permutation procedure (called cross-validation) results in n statistics based an eq. (2.9). Reject the metamodel if <u>any</u> of the n values of the statistic is significant. The proper significance level is based on the Bonferroni inequality, i.e.,the metamodel is rejected if

$$\max_{1 \leq i \leq n} | t_i | > t^{\alpha'} \text{ with } \alpha' = (\alpha_E/n)/2 \qquad (2.10)$$

where α_E is the "experimentwise error rate" and the factor 2 is needed because a two-sided test is appropriate; see Appendix 1 for a discussion on "experimentwise" and "per comparison" error rates.

(iv) If the metamodel postulated in step (i), and estimated ("calibrated") in step (ii), is rejected in step (iii), then several options are available:

- Mechanistic alternative: Add higher-order interactions to the postulated model. This alternative may be inspired by the Taylor series argument; see the discussion around eq. (2.2). In Analysis of Variance high-order interactions are also traditional. This option means that, e.g., (2.5) is augmented with an extra term, namely, $\beta_{123} \cdot z_{i1} \cdot z_{i2} \cdot z_{i3}$. The disadvantages of this option are:
  - o The interpretation of high-order interactions is difficult.
  - o Adding explanatory variables increases $\text{var}(\hat{y})$ and may require extra runs.
- Transformations: For instance, if y denotes waiting time, and x_1 and x_2 denote mean interarrival and service rate, then the transformation $z = x_1/x_2$ yields a better metamodel, as we know from queuing theory. In econometrics the logarithmic transformation is popular since it means that the parameters β can be interpreted as elasticity coefficients. In general, transformations may be based on theory and exploratory data analysis. Simulation without a theoretical or common-sense basis, can never lead to insight! For examples see [9, 14].

- Smaller experimental domain: The Taylor series argument suggests that an approximation may become valid if the domain of the function is reduced. Unfortunately this option also limits the generality of the analyst's conclusions. A small domain is no problem if the objective of the simulation is not to obtain a general understanding but to search for the optimun setting of the quantitative parameters z. In the latter case the analyst can use "Response Surface Methodology" (RSM), explained next.

  RSM runs as follows:

  - o Start in a subdomain, i.e., let the varaiables z not range over the whole experimental area. In this small area the first-order model of eq. (2.3) is applied.
  - o The local first-order model is used to find the direction of steepest ascent for the response.
  - o Repeat fitting first-order models, locally along the path of steepest ascent. When the optimum (possibly a local optimum) is approached, the first-order model is rejected (we would suggest the validation test of eq. 2.9 but this test is no part of traditional RSM). Then a second-order model is fitted, i.e., the model is augmented with interactions as in eq. (2.5) plus "pure quadratic" effects (add $\Sigma\ \beta_{jj}.z_j^2$ to eq. 2.5). Note that adding higher-order effects means that more observations (simulation runs) become necessary; see the "central composite" designs in section 3).
  - o Taking derivatives of the second-order model $\partial/\partial z_j$ and solving $\partial/\partial z_j = 0$ yields the optimum setting of z_j. For literature on RSM see [10, 17].

Metamodels with interactions, pure quadratic effects, or transformed variables are not linear in the independent variables, but these models do remain linear in their parameters $\underline{\beta}$. Hence familiar linear regression analysis applies, i.e., the metamodel is

$$\underline{y} = \underline{Z}.\underline{\beta} + \underline{e} \qquad (2.11)$$

If qualitative factors occur in the metamodel, then the lineair model

of eq. (2.11) still applies. (The model for qualitative factors may be familiar to the reader with a knowledge of Analysis of Variance.) The representation of qualitative factors is first demonstrated for a single qualitative factor z which can assume only two "values" or levels, say, FIFO versus LIFO queuing discipline. These two levels are denoted by -1 and +1 respectively. Hence the response y may be modeled as

$$y_i = \beta_0 + \beta_1 . z_i + e_i \qquad (i = 1,\ldots,n) \tag{2.12}$$

where $z_i = -1$ if the queuing rule is FIFO in simulation run i, etc. Hence the effect of switching from FIFO to LIFO is $2\beta_1$. If $\hat{\beta}_1$ is not significant then the analyst concludes that queuing discipline is unimportant.

If a qualitative factor has more than two levels, then several dummy variables assuming the values 0 or 1, are necessary. For instance, if z assumes three levels, namely level i in run i (i = 1, 2.3), then the model becomes

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 + e_1 \\ y_2 &= \beta_0 + \beta_2 + e_2 \\ y_3 &= \beta_0 + \beta_3 + e_3 \end{aligned} \tag{2.13}$$

with the restriction

$$\beta_1 + \beta_2 + \beta_3 = 0 \tag{2.14}$$

Eq. (2.13) corresponds to eq. (2.11) with

$$\underset{\sim}{Z} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \qquad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \tag{2.15}$$

Although $\underset{\sim}{Z}$ is singular the restriction (2.14) yields unique least squares estimators of $\underset{\sim}{\beta}$; see [10 ,pp.299-301]. Note that eq. (2.13) implies that a change from level 1 to level 2 may produce an effect different

from the effect of changing from level 2 to 3. Note further that only if a factor is quantitative, interpolation and extrapolation are possible.

Usually regression analysis (or equivalently Analysis of Variance) is used to estimate the effects β and to test whether these effects are significantly different from zero. This is a practical approach in the early phase of an investigation when the analyst examines which factors are important; see also the next section. In a later phase, however, he may examine a few remaining system variants, say, ten variants. Then "multiple comparison procedures" (MCP) can compare all system variants with each other (45 comparisons) or can select a subset containing the best variant with prescribed probability, say $1 - \alpha$. Since practitioners show little interest in these procedures we shall not discuss them further but refer to [4,10]. Remember that RSM also investigates systems in detail but then all factors are quantitative.

3. Experimental design

Section 1 has already explained the need for an efficient and systematic way to explore the great many possible system variants (factor-level combinations). Experimental design theory has been widely applied in agricultural and technical experiments. Its application to socio-technical systems is more difficult. However, in a simulation model of whatever system all factors are completely under control so that experimental design theory becomes highly relevant. Because of this complete control traditional design topics like randomization and blocking are unimportant in simulation. The present section contains simple classical designs. Appendix 2 demonstrates that these scientific designs are superior to the "common sense" approach.

(i) Resolution III designs.

Suppose three factors are investigated, so that $k = 3$ in eq. (2.1). Assume further that each factor is studied at only two levels (this assumption may be changed in certain situations discussed later). Hence all 2^3 factor-level combinations might be simulated. However, if the analyst assumes that a first-order metamodel like eq. (2.3) is valid,

he can save 50% of his simulation runs. The explanation is that the first-order model contains only four parameters ($\beta_0, \beta_1, \beta_2, \beta_3$) so that four runs suffice. Which four runs to simulate can be specified by the "tricks" of experimental design (before reading on the reader may try to specify his own selection of four runs). Table 1 shows that the column for the variable x_3 is constructed by multiplying the corresponding elements of the x_1 and x_2 columns.

The symbol x is used in Table 1 and not z as in eq. (2.3), because the experimental design literature gives "normalized" variables x, i.e., variables assuming only the values -1 or +1. The original qualitative variables z (with two levels) are identical to these x; see the discussion around eq. (2.12). The original quantitative variable z_j (assuming values between L_j and H_j) is derived from x_j through the lineair transformation

$$z_{ij} = a_j \cdot x_{ij} + b_j \qquad (3.1$$

with

$$a_j = (H_j - L_j)/2 \qquad (3.2)$$

$$b_j = (H_j + L_j)/2 \qquad (3.3)$$

In general, Resolution III (R III) designs assume that a first-order model with q = k+1 parameters holds, and permits the estimation of these k + 1 parameters in only n = [k + 1] runs (where [x] means that x is rounded upwards to the next multiple of 4, e.g., if k is 12 or 15 then n becomes 16). For high values of k dramatic savings result, e.g., if k is 7 then n is 8 while simulating all combinations would require $2^7 = 128$ runs; see the upper part of Table 2 where $x_{4i} = x_{1i} \cdot x_{2i}$ or in short-hand notation $\underset{\sim}{4} = \underset{\sim}{1}.\underset{\sim}{2}$; likewise $\underset{\sim}{5} = \underset{\sim}{1}.\underset{\sim}{3}$, $\underset{\sim}{6} = \underset{\sim}{2}.\underset{\sim}{3}$ and $\underset{\sim}{7} = \underset{\sim}{1}.\underset{\sim}{2}.\underset{\sim}{3}$.

The designs demonstrated in Table 1 and (the upper part of) Table 2 are a special subclass of R-III designs, namely 2^{k-p} designs (p denotes the fraction of the 2^k combinations not executed). If n is

Table 1

Experimental Design for Three Factors

Combination i	x_{1i}	x_{2i}	$x_{3i} (= x_{1i} \cdot x_{2i})$
1	+1	+1	+1
2	-1	+1	-1
3	+1	-1	-1
4	-1	-1	+1

Table 2

Experimental Design for Seven Factors

Combination	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	-	-	-	+	+	+	-
2	+	-	-	-	-	+	+
3	-	+	-	-	+	-	+
4	+	+	-	+	-	-	-
5	-	-	+	+	-	-	+
6	+	1	+	-	+	-	-
7	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+
9	+	+	+	-	-	-	+
10	-	+	+	+	+	-	-
11	+	-	+	+	-	+	-
12	-	-	+	-	+	+	+
13	+	+	-	-	+	+	-
14	-	+	-	+	-	+	+
15	+	-	-	+	+	-	+
16	-	-	-	-	-	-	-

not a power of 2 (but still a multiple of 4), e.g. n = 12, then the trick of identifying first-order effects with interactions does not work. Instead a table of so called Plackett-Burman designs has to be consulted which specifies $\underset{\sim}{X}$; see [10] for such a table.

Advantages of R-III designs are; see also [10, p 370]:

- Small number of runs: $n \geq (k+1)$ but $n << 2^k$.
- Maximum accuracy (?): It can be shown that under the classical statistical assumptions concerning $\underset{\sim}{e}$ (independent errors with constant variances) the variances of the OLS estimators of $\underset{\sim}{\beta}$ are minimized if $\underset{\sim}{X}$ is orthogonal. R-III designs are orthogonal indeed, as Tables 1 and 2 demonstrate. If $\underset{\sim}{\Omega}_e \neq \sigma^2 . \underset{\sim}{I}$ then it is unknown whether R-III designs yield optimal results, but these designs certainly provide non-singular $\underset{\sim}{X}$. Moreover, in simulation it would be possible to run a factor-level combination so long that all combinations have the same (estimated) variance; however, we do not know applications of this approach.
- Model validation: If $n > k + 1$ the cross-validation approach of the previous section is possible. This condition is met if $k + 1$ is not exactly a multiple of four. If $k + 1$ is exactly a multiple of four one or more extra runs can be added to the R-III design to validate the first-order model. If a factor is quantitative, then this extra run may correspond to the "central" value $x = 0$, so that the analyst can check whether pure quadratic effects are zero indeed. If the analyst feels from the start that a first-order metamodel may very well be too simple, then the next type of designs can be useful.

(ii) Higher resolution designs.

If more than $k + 1$ combinations are simulated then estimators of higher-order effects become possible in principle. Experimental design theory provides a number of design types. The following types remain quite simple.

- Resolution V (R-V) designs

  By definition R-V designs permit estimation of all two-factor interactions $\beta_{jj'}$; see eq. (2.5). Unfortunately the total number of parameters q increases drastically with k, since there are $k.(k-1)/2$

two-factor interactions plus k first-order effects and one grand-mean. There are several types of R-V designs, e.g., 2^{k-p} designs of course p is lower than in R-III designs). Designs where n is not higher than q have been proposed by Rechtschaffner; see [10], but we do not know any practical applications of Rechtschaffner's designs.

- Resolution IV (R-IV) designs.

If all factors are qualitative, then R-IV designs permit the unbiased estimation of all first-order effects even if two-factor interactions are important; at the same time these designs provide estimators of certain sums of interactions. For instance, if k = 7 then the sixteen runs of Table 2 provide unbiased estimators of β_j (j = 1,...,7) plus estimators of the sums $\beta_{24} + \beta_{35} + \beta_{67}$, $\beta_{14} + \beta_{36} + \beta_{57}$, etc. (assuming no interactions among more than two factors are important). If all sums of interactions are non-significant then the first-order model has been estimated and validated. Otherwise additional experimentation is necessary but the R-IV design may suggest the elimination of one or more factors in future experimentation, namely if one or more first-order effects are non-significant. Applications of R-IV can be found in [10].

Technically R-IV designs can be constructed very simple: once a R-III design is available, just duplicate this design with reversed signs,i.e., in the lower part of Table 2 $x_{ij} = -x_{j'j}$ (i' = 1,...,7; i = 8,...,16). So a R-IV design requires twice as many runs as a R-III design. Note that once a design has been selected, cross-products $x_j x_{j'}$, are also fixed.

- Response Surface Methodology designs.

If all factors are quantitative then RSM may be applied; see Section 2. As long as the optimum region is not approached first-order models guide the search so that R-III designs can be used. In the optimum region a second-degree polynomium is used, including k pure quadratic effects β_{jj}. Then "central composite" designs can be applied, i.e., the R-IV design is augmented with 2k "axial" points: $x_j = -c$ and $x_j = +c$ respectively while $x_{j'} = 0 (j \neq j')$. Moreover the central point ($\forall j \; x_j = 0$) is replicated a few times. Under specific statistical assumptions like constant variances the central composite design has certain optimum properties like minimum bias caused by possible third-order effects.

- Other designs.

In a simulation study of the Rotterdam harbor six factors were studied; not all interactions were thought to be equally important; actually only the following interactions were expected to be interesting: β_{12}, β_{13}, β_{23}, β_{24}, β_{25} and β_{26}. Hence the total number of parameters in the postulated metamodel was q = 13. Using the "tricks" of experimental design these 13 parameters were estimated in only 16 runs (trick: $\underset{\sim}{1} = \underset{\sim}{5}.\underset{\sim}{6}$ and $\underset{\sim}{3} = \underset{\sim}{4}.\underset{\sim}{5}$). Of course some extra runs were used to validate the resulting metamodel; see [14].

In general experimental design theory clearly shows how estimators of effects are biased by other effects. For instance, in the last example $\hat{\beta}_1$ is not biased by β_{12} (suspected to be an important interaction) but β_1 is biased by β_{56} (but this interaction is assumed to be unimportant). Another example is Table 1 where $\hat{\beta}_1$ is biased by β_{23} but not by β_2 or β_3. Clearly the choice of a design is based on a postulated metamodel.

(iii) Screening designs.

For pedagogical reasons screening designs are discussed after the other types of designs. In practical applications screening designs, however, play a role in the very first stage of experimentation. Screening designs are namely meant to investigate a great many factors, e.g., k = 1000. So when the analyst has constructed a first version of his simulation model, this model contains many factors of unknown importance but - hopefully - only relatively few (say, k' << k) factors are really important. (Otherwise everything depends on everything else and a scientific explanation breaks down.) The analyst may then use screening designs to detect the k' important factors. (After this screening or pilot phase the important factors can be investigated in detail, applying the designs discussed earlier).

Until now practical applications of screening designs have been rare. The reason may be that academic problems are usually small-scale so that screening is irrelevant. Practical problems are often large-scale but then the necessary statistical know-how is usually missing. In practice one type of screening design is sometimes applied,

namely random designs. This type is easily constructed: sample the factor-level combinations randomly. The price paid for this simplicity is that control over the factor levels is relinguished [10]. We recommend that the analyst uses his control over the simulation axperiment as follows.

Group the k original factors into a much smaller number of groups. Under mild assumptions discussed below, a group will be significant if and only if that group contains one or more important original factors. Since in the pilot phase many individual factors are unimportant, many groups are unimportant. Future experimentation can then ignore all individual factors within non-significant groups. Let us consider this procedure in more detail.

First suppose a first-order metamodel in the original factors x is valid. (In the pilot phase factors are treated as qualitative, on/off variables; with qualitative factors the assumption of a first-order model can be easily relaxed; see below.) Moreover assume that the signs of the effects are known, e.g., if the number of service stations is really important then it has a negative effect on the waiting time (this assumption is discussed below). Hence the levels of the factors x can be so defined that possible effects β are positive or zero but certainly not negative: $\beta_j \geq 0$ $(j = 1,\ldots,k)$. Now a "group factor", say w_1, has the level -1 if all its members x_g are at their low-level -1 ($g = 1,\ldots,G$ with group size G). And when $w_1 = +1$ then $x_g = +1$. Next consider an example with $k = 100$ and $G = 50$ so that only two group factors w_1 and w_2 result. These two group factors can be investigated in a classical 2^2 design. Let all effects β be zero except for β_1 and β_2. Then Table 3 results. Hence the effect γ_1 of group factor w_1 is estimated by

$$\hat{\gamma}_1 = \sum_{i=1}^{n} w_{i1} \cdot y_i / n \qquad (3.4)$$

with expected value

$$E(\hat{\gamma}_1) = (4\beta_1 + 4\beta_2)/4 = \beta_1 + \beta_2 \qquad (3.5)$$

whereas analogous reasoning yields $E(\hat{\gamma}_2) = 0$. Hence after a screening phase of four runs all fifty individual factors within group factor 2 can be eliminated. In the next stage group factor 1 can be group-screened further, etc.

How restrictive are the assumptions of group screening? If a few individual factors have unknown signs then these factors only can be placed in groups of size one. Alternatively the analyst may rely on the low probability of several important factors x being placed within the same group and moreover having opposite signs of the same magnitude so that their effects cancel out. The other assumption, a first-order metamodel, can be relaxed as follows. If the metamodel is postulated to include two-factor interactions, then the group factors should be investigated, not in a R-III design, but in a R-IV design. (For instance, Table 3 is actually a R-IV design; if $\beta_{12} \neq 0$ then add $+\beta_{12}$ to all responses; compute $\hat{\gamma}$ as in eq. (3.4); find that in $E(\hat{\gamma}_1)$ β_{12} vanishes; likewise β_{12} vanishes in $E(\hat{\gamma}_2)$.) For details see [10].

Practical applications of group screening are rare. We can mention only two references [18,19], one on the simulation of a strategic airlift and one on a computer system simulation. In an unpublished Monte Carlo experiment we found that group screening indeed detected the seven important factors among the 88 original factors.

Concluding the discussion of the strategic issues we emphasize that statistical techniques alone cannot solve the analyst's problem. The model specification and the hypotheses are not provided by statistical theory. Moreover the statistical techniques do not specify which α value to take, etc. Finally, statistical techniques are based on statistical assumptions like normality so that the analyst has to use his judgement in considering the sensitivity of the technique to these assumptions (robustness issue). Nevertheless statistical techniques can drastically improve "common sense" approaches.

Table 3

Group screening example

Run i	Group factor w_1	w_2	Individual factor x_1	x_2	x_3 ...	x_{50}	x_{51} ...	x_{100}	Response $E(y)-\beta_0$
1	-	-	-	-	- ...	-	- ...	-	$-\beta_1$ $-\beta_2$
2	-	+	-	-	- ...	-	+ ...	+	$-\beta_1$ $-\beta_2$
3	+	-	+	+	+ ...	+	- ...	-	β_1 $+\beta_2$
4	+	+	+	+	+ ...	+	+ ...	+	β_1 $+\beta_2$

4. Runlength and confidence intervals.

As we mentioned in the Introduction runlength determination is a tactical issue. Much has been published on this problem; see [2,20]. Many publications ignore the fact that in practical situations - as opposed to academic simulations - the runlength issue can be solved with elementary statistics rather than with complicated theory on stochastic processes. The explanation is that in practice simulations are usually "terminating", i.e., the run is stopped when a specific event occurs. Some examples are:

- A queuing system such as a bank closes at 5 P.M. (critical event: 5 P.M.). The next run start on a new day, in the empty state.
- A corporate simulation investigating how profit reacts to different policies, stops when the planning horizon of three months is reached. The runs start from the most recent situation of the company.
- When studying maintenance strategies the simulation starts with a new piece of equipment and stops when the equipment breaks down.
- Queuing systems are often simulated to see if they can handle peak traffic. The run starts before the peak and stops when the peak is over. This example is more difficult than the preceding three examples since the start condition and the critical event are not sharply defined: when exactly is the rush hour? We emphasize that practioners often do not realize that their simulation should not be run indefinitely long with the peak hour traffic intensity. We do not know any publications where the runlenght problem in peak systems is examined in detail.

In academic studies simulation is often used to estimate the expected response in the steady state. Then we have non-terminating simulation (see later). In the above examples, however, start-up (transient) and end effects are part of the response. One run from the start condition until the critical event yields one observation y. More accurate estimates are possible, repeating the run with different random number streams. The statistical analysis requires only elementary statistics for the standard error after n runs is

$$s = \{ \sum_{i=1}^{n} (y_i - \bar{y})^2 / (n - 1) \}^{1/2} \qquad (4.1)$$

The $(1 - \alpha)$ confidence interval can be based on the Student t statistic (y is often approximately normal as explained by the central limit theorem or a related theorem; moreover t is a robust statistic):

$$P\ (\ E(y) \in \bar{y} \pm t^{\alpha/2}_{n-1}\ .\ s/\sqrt{n}) = 1 - \alpha \qquad (4.2)$$

If the confidence interval is found to be too wide, additional runs can be generated. It seems straightforward to derive n_c, the total number of runs necessary to obtain a confidence interval with lenght c:

$$n_c = (t^{\alpha/2}_{n_c-1}\ .\ s\ /\ c)^2 \qquad (4.3)$$

However, a complication is that n in eq. (4.2) is deterministic but n_c in eq. (4.3) has become stochastic. Fortunately it can be proven that eq. (4.3) gives satisfactory results, if y is indeed independent and normally distributed. In practice some care is needed since if y is not normal then y and s become dependent, so that the coverage of the confidence interval may deviate from $1 - \alpha$. Anyhow eq. (4.3) is superior to the "practical" method of executing so many runs that, say, the third digit of the averaged output does not change.

In the literature most attention is focussed on steady-state behavior (even though in practice most systems are terminating). Then the analyst has several options:

(i) Replicated runs.

The analyst may try to use the same technique as used with terminating systems, i.e., he may replicate the simulation run n times, each run with different random number seeds. The advantage is that these n runs are independent. Unfortunately in steady-state problems there is an initialization problem. Whereas in terminating systems the transient observations are part of the relevant output, in steady-state situations start-up observations create bias. Hence two alternatives are available:

- Retain the transient phase.
  Though the start-up phase creates bias, this phase does contain information. Hence it is very well possible that Mean Squared Error (MSE) is minimized if all observations are used. However, minimizing MSE may

yield invalid confidence intervals.

- Eliminate the transient phase. But two practical problems remain:
  - o How can the duration of the transient phase be determined? The practical solution is to make graphs and see whether start-up effects seem to have disappeared. We do not know practical statistical techniques for this problem.
  - o With n replications n transient phases are thrown away. To avoid this waste, several other approaches have been developed.

Instead of replicating runs the analyst can continue the simulation run a very long time (remember no critical event terminates the run). This single run comprises many individual observations, say individual waiting times w. Unfortunately these individual observations are autocorrelated. In most systems the autocorrelations are positive, e.g., if a customer has to wait long, then the chance of a long waiting time for the next customer increases. Consider the individual observations w_t with autocorrelation ρ_j between w_t and w_{t+j} and with variance $\text{var}(w_t) \equiv \sigma_w^2 \equiv \rho_0$ (the w_t form a stationary stochastic process since ρ and σ_w^2 do not depend on t). Then the variance of the average $\bar{w}$ is:

$$\text{var}(\bar{w}) = \{1 + 2 \sum_{j=1}^{m} (1 - \frac{j}{m}) \cdot \rho_j \} \cdot \sigma_w^2/m \qquad (4.4)$$

Applying elementary statistical techniques assuming independence among the w, would be very misleading. Assuming independence eq. (4.4) reduces to the familiar expression $\text{var}(\bar{w}) = \sigma_w^2/m$. The positive autocorrelations inflate the variance of the average. For instance, in a Markovian single-server system with a utilization of 50% the inflation factor (the curly brackets in eq. 4.4) becomes 10; for a traffic load of 90% this factor is as high as 360. Therefore option (i) is augmented with several more options.

(ii) Estimate the autocorrelation structure

The analyst may estimate the autocorrelations $\rho_{j'}$ in eq. (4.4) ($j' = 0, 1, \ldots$). Spectral analysis is a technique for estimating not the $\rho_{j'}$ themselves but their Fourier transformations [2, 6]. Unfortunately there are a number of technical problems, e.g., at which value of m in eq. (4.4) should the analysis stop? Moreover for most practioners spectral

analysis is too sophisticated. Of course, practioners may use an appropriate statistical package as a black box, but such an approach certainly has its disadvantages.

(iii) Cut the run into nearly independent subruns.
In practice the prolonged run is often cut into, say, S subruns of lenght L = m/S, and the subrun averages $\bar{w}_s$ are treated as S independent observations to which eqs. (4.1) through (4.3) are applied, replacing y_i in those equations by $\bar{w}_s$, etc. The underlying idea is that although the first few observations of a subrun still depend on the last few observations of the preceding subrun, the subrun averages are practically speaking independent, if at least the subruns are long enough. Obviously the stronger the autocorrelations among the individual observations, the longer the subruns should be. Therefore the analyst may start with some intuitively selected subrun lenght L_0, compute the resulting subrun averages; and statistically test the independence of these averages (e.g. through the Von Neumann test). If the selected subrun length creates significant dependence, then the subrun length is increased, etc. [2].

In practice the subrun approach is often followed, but without testing the appropriateness of the intuitively selected subrun size. Relying on intuition only is dangerous, since analytical results for simple queuing systems have demonstrated that individual observations on heavy-traffic systems remain autocorrelated over surprisingly long intervals; see also the comment below eq. (4.4). On the other side, a too conservative subrun size means that few subrun averages remain so that the resulting confidence interval becomes less stable. A practical heuristic is to take ten to twenty subruns to estimate the confidence interval for the mean (but to take at least 100 subruns to test their independence); [15].

(iv) Renewal analysis.
Stochastic systems may have a renewal or regenerative property. For instance, a queuing system with utilization less than 100%, becomes idle now and then; the next history is then independent of the past

history if the arrival process is Poisson (memoryless). So"subruns" (or cycles, tours) can then be defined such that a subrun starts as soon as a customer arrives into an empty system. In contrast to option (iii) renewal analysis creates perfectly independent subruns; the S subruns have different lenghts L_i ($i = 1,...S$) depending on when the subruns return to the renewal state.

In general, any Markov system has the renewal property. A practical problem is that the renewal state may occur so infrequently that only few (independent) subruns result. For instance, if the system has heavy traffic then the empty state occurs rarely. In another situation the system may have so many states that the realization of one particular state occurs rarely. A practical remedy is to define an approximate renewal state, e.g., let a new subrun start as soon as the system is "nearly" empty [3].

From a statistical viewpoint it is interesting to note that the renewal analysis of stochastic processes involves ratio estimation. For instance, to estimate the expected steady-state waiting time, say μ, subrun s with lenght L_s yields the subrun's total waiting time Y_s computed from the individual waiting times $w_{s\ell}$:

$$Y_s = \sum_{\ell=1}^{L_i} w_{s\ell} \quad (s = 1,...,S) \tag{4.5}$$

Then the traditional estimator of μ is rewritten as

$$\bar{w} = \sum_{j=1}^{m} w_j / m = \sum_{s=1}^{S} Y_s / \sum_{s=1}^{S} L_s = \bar{Y} / \bar{L} \tag{4.6}$$

The following $1-\alpha$ confidence interval can be derived, applying the central limit theorem:

$$\bar{w} \pm z^{\alpha/2} \cdot (\hat{\sigma} / \sqrt{S}) / \bar{L} \tag{4.7}$$

with

$$\hat{\sigma}^2 = \hat{\sigma}_{11} - 2\,\bar{w} \cdot \hat{\sigma}_{12} + (\bar{w})^2 \cdot \hat{\sigma}_{22} \tag{4.8}$$

where $\hat{\sigma}_{11}$, $\hat{\sigma}_{22}$ and $\hat{\sigma}_{12}$ are the usual estimators of var (Y), var (L) and cov (Y,L). However, the estimator of eq. (4.6) is only one of the possible estimators of μ. For it can be shown that

$$\mu = E(Y) / E(L) \qquad (4.9)$$

and to estimate this ratio several point estimators and confidence intervals (including jackknifing) an available; see [1,2]. These procedures can be learned by practitioners without much trouble, and applications have indeed started to appear. In the mean time research on renewal analysis continues, e.g., recently regression and graphical techniques have been developed to detect and diminish small-sample bias and nonnormality of renewal estimators; [5].

Note that the initialization problem is severe in option (i). But in options (ii) and (iii) transient observations also have to be removed. Option (iv), however, has no initialization problem since observations can be collected immediately when the simulation starts in the renewal state, e.g., the idle state.

The discussion of this section has concentrated on the estimation of the expected value of the response variable, say E(y). This emphasis is in line with most publications. In practice the user is often interested not only in the mean but also in quantiles, say y_p with $P(y < y_p) = 1 - p$. Sometimes y_p is fixed, e.g., what is the chance that customers have to wait longer than two seconds. Then the analyst has to estimate the probability 1 - p and he may introduce the binary variable, say

$$\begin{aligned} v &= 0 \text{ if } y > y_p \\ &= 1 \text{ if } y \leq y_p \end{aligned} \qquad (4.10)$$

It is also possible that not y_p but p is fixed, e.g., estimate the response time not exceeded by 95% of the customers. Then y_p can be estimated, arranging the observations y in increasing order, so-called order statistics. Publications on quantile estimation are extremely rare (except for some recent working papers; see also [8]).

5. Variance reduction techniques (VRT).

Many VRT are available [10] but most techniques are too sophisticated for practical use. Moreover computer time can be saved by other means, e.g. more efficient sampling of input variables, better simulation languages. Therefore this section emphasizes a technique that has already been used by practitioners (but often without proper statistical care).

(i) Common random numbers.

Several systems variants may be simulated using the same random number seed. This technique has great intuitive appeal since it means that, e.g., queuing system variants are simulated with the same customers. We add that in complicated systems care should be taken to synchronize the random number streams in different system variants, e.g., arrival and service processes may use separate streams. An issue overlooked by most practitioners is the statistical analysis of the simulation output which becomes more complicated when responses are dependent. When estimating the regression metamodel's parameters $\underset{\sim}{\beta}$ Generalized Least Squares (GLS) may be considered, so that in eqs. (2.6) and (2.7) $\underset{\sim}{D}$ is replaced by a covariance matrix $\underset{\sim}{\Omega}_y$ with positive elements off the main diagonal. When using OLS the standard errors of the estimated $\underset{\sim}{\beta}$ follow not from the "classical" formula (replace $\underset{\sim}{D}$ by $\sigma^2.\underset{\sim}{I}$ in eq. 2.7), but from

$$\underset{\sim}{\Omega}_{\underset{\sim}{\beta}} = (\underset{\sim}{X}'.\underset{\sim}{X})^{-1} \; . \; \underset{\sim}{X}' . \; \underset{\sim}{\Omega}_y . \; \underset{\sim}{X} \; (\underset{\sim}{X}'.\underset{\sim}{X})^{-1} \qquad (5.1)$$

Classical experimental designs have certain statistical optimality qualities like minimum variance, if the responses are independent with constant variances. If this condition does not hold, the designs are not known to be optimal (but they are still superior to the "common sense" approach). If the use of common random numbers does create the desired positive correlations, then common random numbers applied in an experimental design do create more accurate estimated effects; see [21].

(ii) Other simple VRT.

There are more VRT, some techniques only slightly more complicated than using common random numbers. Nevertheless even a minor complication seems enough to preclude practical use. Moreover it cannot be guaranteed that these techniques yield substantial variance reduction. The reader who wants to consider VRT, is advised to start by looking at the following techniques [10]:

- "Antithetic variates": this is a very simple technique which does not complicate the statistical analysis. This technique is relevant if more than a single run is generated per factor-level combination.
- "Control variates" or "regression sampling": Compared to antithetic variates this technique is more complicated.
- "Importance sampling" and "virtual measures": this technique is really complicated and should only be studied when extremely much computer time is needed per factor-level combination, in order to estimate very rare events like "excessive" waiting times, and nuclear disasters [7].

6. Conclusions.

Since simulation means experimentation (albeit with a model instead of a real-world system), statistical analysis and design techniques are necessary. The following strategic and tactical issues have been discussed.

Strategic issues:

(i) In order to generalize and interpret simulation output a metamodel is useful for which regression analysis can be applied. This metamodel is a first-order or second-order approximation to the intricate simulation program. The metamodel can be tested for its adequacy. Weighted Least Squares is recommended (unless common random numbers are used).

(ii) To decide which factor-level combinations (system variants) should actually be simulated, experimental designs can be applied.These designs provide more accurate estimators and enable estimation of interactions, often saving simulation runs at the same time.

(iii) In the preliminary phase of simulation experimentation screening designs can be used to investigate, say, hundreds of factors in relatively few runs.

Tactical issues:

(i) The analyst should be aware of the distinction between terminating and non-terminating systems.

(ii) For terminating systems replication results in a simple confidence interval procedure. Initialization is no major problem.

(iii) For steady-state estimation a number of procedures are available, each with its advantages and disadvantages.

(iv) Variance reduction techniques are available but only the common random numbers technique is popular. The latter technique slightly complicates the statistical analysis.

Appendix 1 The experimentwise error rate.

In many experiments a number of responses are investigated, e.g., both waiting time and server utilization. Moreover a single response variable may be characterized by several measures, e.g. its mean and its 95% quantile. Suppose that for N variables a $1 - \alpha$ confidence interval is derived, e.g., an interval as in eq. (4.2). For illustration purposes assume that $N = 20$ and $\alpha = 0.10$. Consequently we expect that two intervals will fail to contain the true values. To ensure - with probability $1 - \alpha$ - that all twenty intervals contain the corresponding true values, the individual intervals must be made wider. More exactly, let $1 - \alpha_E$ denote the probability that all statements made in the analysis of the experiment, are correct. Let $1 - \alpha_C$ be the probability that an individual statement is correct; α_C is the "per comparison" error rate used in elementary statistics, e.g., in eq. (4.2) $\alpha_C = \alpha$. If the individual statements are independent then

$$1 - \alpha_E = (1 - \alpha_C)^N \qquad \text{(A1.1)}$$

If the statements are dependent, then the Bonferroni inequality applies:

$$\alpha_E \leq N \cdot \alpha_C \qquad \text{(A1.2)}$$

Actually different values for α_C may be used, say, α_{Ci} and then the Bonferroni inequality yields

$$\alpha_E \leq \sum_{i=1}^{N} \alpha_{Ci} \qquad \text{(A1.3)}$$

For simultaneous inferences or for multivariate responses the Bonferroni inequality provides an extremely simple technique, which is often superior to more sophisticated techniques; [10,16] . Note that the experimentwise error rate is controlled at the expense of wider confidence intervals (i.e. decreased power of the corresponding hypothesis tests).

Appendix 2: One-factor-at-a-time versus factorial experimentation.

In practice simulation models are usually explored, changing one factor at a time. Sometimes this method is even heralded as "the" scientific method. A simple example shows that factorial methods yield the same type of estimators but make these estimators more accurate; moreover factorial experiments provide some extra estimators. For illustration purposes assume there are only two factors. Then as Table 4 shows the one-factor-at-a-time method yields

$$\hat{\beta}_1 = y_1 - y_2 \;,\; \hat{\beta}_2 = y_3 - y_4 \qquad \text{(A2.1)}$$

with var $(\hat{\beta}) = 2\,\sigma^2$ (assuming constant variances σ^2). The factorial experiment yields responses y' so that

$$\hat{\beta}_1 = (y_1' + y_3')/\,2 - (y_2' + y_4')/\,2 \qquad \text{(A2.2)}$$

$$\hat{\beta}_2 = (y_1' + y_2')/\,2 - (y_3' + y_4')/\,2 \qquad \text{(A2.3)}$$

with var $(\hat{\beta}) = \sigma^2$. Moreover it becomes possible to estimate the interaction β_{12} (construct a column $x_{1i}.x_{2i}$):

$$\hat{\beta}_{12} = (y_1' - y_2' - y_3' + y_4')/\,2 \qquad \text{(A2.4)}$$

If many factors are to be examined, then the "practical" method requires 2 k runs. Each estimated first-order effect has variance $2\,\sigma^2$. A R-III design would requiring only n = k+1 runs and yield estimators with variance $4\,\sigma^2/n$. If (roughly) the same number of runs as with the

Table 4

One-factor-at-a-time versus factorial experimentation

Run	One-factor-at-a-time		Response	Factorial		Response
i	x_1	x_2	y	x_1	x_2	y'
1	+1	0	y_1	+	+	y_1'
2	-1	0	y_2	-	+	y_2'
3	0	+1	y_3	+	-	y_3'
4	0	-1	y_4	-	-	y_4'

"practical" method are executed then a R-IV design becomes possible with all the additional information provided by such a design.

References.

1. CRANE, A. and J. LEMOINE, AN INTRODUCTION TO THE REGENERATIVE METHOD FOR SIMULATION ANALYSIS. Springer-Verlag, Berlin, 1977.

2. FISHMAN, G.S., PRINCIPLES OF DISCRETE EVENT SIMULATION. Wiley-Interscience, New York, 1978.

3. GUNTHER, F.L. and R.W. WOLFF, The almost regenerative method for stochastic system simulations. OPERATIONS RESEARCH, 28, no. 2, March-April 1980, pp.375-386.

4. GUPTA, S.S. and S. PANCHAPAKESAN, MULTIPLE DECISION PROCEDURES: THEORY AND METHODOLOGY OF SELECTING AND RANKING POPULATIONS. John Wiley & Sons, Inc., New York, 1979.

5. HEIDELBERGER, P. and P.A.W. LEWIS, Regression - adjusted estimates for regenerative simulations, with graphics. COMMUNICATIONS ACM, 24, no. 4, April 1981, pp. 260-273.

6. HEIDELBERGER, P. and P.D. WELCH, A spectral method for confidence interval generation and run length control in simulations. COMMUNICATIONS ACM, 24, no. 4, April 1981, pp. 233-245.

7. HOPMANS, A.C.M. and J.P.C. KLEIJNEN, Importance sampling in systems simulation: a practical failure? MATHEMATICS AND COMPUTERS IN SIMULATION, 21, 1979, pp. 209-220.

8. IGLEHART, D.L., Simulating stable stochastic systems, VI: quantile estimation. JOURNAL ASSOCIATION COMPUTING MACHINERY, 23, no. 2, April 1976, pp. 347-360.

9. KEYZER, F., J. KLEIJNEN, E. MULLENDERS and A. VAN REEKEN, SMALL-JOBS-FIRST; A COMBINED QUEUEING, SIMULATION, AND REGRESSION ANALYSIS. Department of Business and Economics, Tilburg University, Tilburg (Neth.), Dec. 1980. (Submitted for publication.)

10. KLEIJNEN, J.P.C., STATISTICAL TECHNIQUES IN SIMULATION (In two volumes.) Marcel Dekker, Inc., New York, 1974/1975. (Russian translation: Publishing House "Statistics", Moscow, 1978.)

11. KLEIJNEN, J.P.C., The role of statistical methodology in simulation. In: METHODOLOGY IN SYSTEMS MODELING AND SIMULATION, edited by B. ZEIGLER, M.S. ELZAS, G.J. KLIR and T.I. ÖREN, North-Holland Publishing Company, Amsterdam, 1979.

12. KLEIJNEN, J.P.C., CROSS-VALIDATION USING THE t STATISTIC. Department of Business and Economics, Tilburg University, Tilburg (Neth.), August 1981.

13. KLEIJNEN, J.P.C., R. BRENT and R. BROUWERS, Small-sample behavior of weighted least squares in experimental design applications. COMMUNICATIONS IN STATISTICS, SIMULATION AND COMPUTATION, B 10, no. 3, 1981, pp.303-313.

14. KLEIJNEN, J.P.C., A.J. van den BURG and R. T. van der HAM, Generalization of simulation results: practicality of statistical methods. EUROPEAN JOURNAL OPERATIONAL RESEARCH, 3, 1979, pp.50-64.

15. KLEIJNEN, J.P.C., R. van der VEN and B. SANDERS, Testing independence of simulation subruns: a note on the power of the Von Neumann test. EUROPEAN JOURNAL OPERATIONAL RESEARCH, forthcoming.

16. MILLER, R.G., SIMULTANEOUS STATISTICAL INFERENCE, McGraw-Hill Book Company, New York, 1966.

17. MONTGOMERY, D. C. and V. M. BETTENCOURT, Multiple response surface methods in computer simulation. SIMULATION, 29, no. 4, Oct. 1977, pp.113-121.

18. NOLAN, R. L. and R. MASTROBERTI, Productivity estimates of the strategic airlift system by the use of simulation. NAVAL RESEARCH LOGISTICS QUARTERLY, 19, 1972, pp.737-752.

19. SCHATZOFF, M. and C. C. TILLMAN, Design of experiments in simulation validation. IBM JOURNAL RESEARCH AND DEVELOPMENT, 19, no. 3 May 1975, pp. 252-262.

20.SCHRIBER, T.J. and R.W. ANDREWS, A conceptual framework for research in the analysis of simulation output. COMMUNICATIONS ACM, 24, no. 4, April 1981, pp. 218-232.

21.SCHRUBEN, L.W. and B.H. MARJOLIN, Pseudorandom number assignment in statistically designed simulation and distrubution sampling experiments. JOURNAL AMERICAN STATISTICAL ASSOCIATION, 73, no. 363, Sept. 1979, pp. 504-525.

22.WEEKS, J.K. and J.S. FRYER, A methodology for aasigning minimum cost due - dates. MANAGEMENT SCIENCE, 23, no. 8, April 1977, pp. 872-881.