Sensitivity analysis and optimization of simulation experiments, including case studies

Kleijnen, J.P.C.

Published in:
EUROSIM Simulation News Europe

Publication date:
1992

Sensitivity Analysis and Optimization of Simulation Experiments, including Case Studies

Jack P.C. Kleijnen
Professor of Simulation and Information Systems
School of Management and Economics
Tilburg University (Katholieke Universiteit Brabant)
5000 LE Tilburg, Netherlands
E-mail: kleijnen@kub.nl, Fax: 013-663072

This essay was written at the request of the editors of 'EUROSIM - Simulation News Europe', which is published on behalf of the Federation of European Simulation Societies.

Abstract

This tutorial gives a survey of what-if analysis and optimization of simulation models, using statistical techniques for the design and analysis of experiments with these models. The simulation models may be deterministic or random. The statistical analysis uses regression (meta)models and Least Squares. The design uses classic experimental designs such as 2^{k-p} factorials, which are both efficient and effective. If there are hundreds of simulation inputs, then special techniques such as group screening and sequential bifurcation may be used. This overview emphasizes applications.

1. Introduction

By definition, simulation involves experimentation, namely with the *model* of a real system. Experimentation requires an appropriate *statistical design and analysis*. Pertinent questions are: which combinations of simulation inputs should be simulated, and how can the resulting output be analyzed? These questions arise in both random and deterministic simulations. Mathematical statistics can be applied to solve these problems, also in deterministic simulation; see Kleijnen and Van Groenendaal (1992, p. 154). In this tutorial I emphasize simple 'academic' examples and practical case studies; for details on statistical procedures I refer to the literature. To save space I refer to a recent introductory textbook, namely Kleijnen and Van Groenendaal (1992), which contains many references for further study. Note that sensitivity analysis and optimization are also addressed in *model validation*, *what-if analysis*, and *goal seeking*.

2. Regression Metamodels

It is good practice to present the results of simulation experiments in the form of graphs. Consider an 'academic' example, namely a single-server queuing system. Its average simulated waiting time (say) y may be displayed as a function of mean service time (say) x. If the simulation experiment is restricted to a small range of values for x, then $y = a + b\,x$ is probably an adequate approximation for the input/output behavior of the simulation model. An alternative presentation replaces x by 1/x (service rate). If the 'domain of experimentation' is wider or if the queuing system is nearly saturated (high traffic loads lead to 'exploding' waiting times), then an alternative display uses $y = a + b\,x + c\,x^2$: second-order approximation (based on Taylor series expansion). One more alternative, however, uses logarithmic paper: $\ln(y) = a + b \ln(x)$ (or $y = c\,x^b$ with $a = \ln c$). Moreover, even this simple simulation model has more than one input: besides mean service time there is mean interarrival time. One possibility is to superimpose upon the graph discussed so far, another graph for a different arrival rate. The latter graph lies above the former graph if the new arrival rate is higher. If there is no *interaction* between arrival and service rates, then (by definition) the two graphs are parallel. Queuing theory suggests that it is better to present average waiting time as a function of the traffic load (arrival rate/service rate), instead of the two individual inputs. This example may be extended: present one graph for one server, superimpose the graph for two servers, etc.; see Kleijnen and Van Groenendaal (1992, pp.159-162).

Regression analysis formalizes this graphic presentation of simulation results: least squares is used to fit a regression model (like $y = a + b\,x$) to the simulation input and output data; the regression model is generalized to include several simulation inputs ('multiple' regression analysis); and the importance of these inputs can be tested, under certain statistical assumptions. I call the regression model a *metamodel* because it models the input/output behavior of the underlying simulation model; the latter model is treated as a black box.

Before systems analysts start experimenting with a simulation model, they have accumulated *prior knowledge* about the real system: they may have observed the real system, tried different simulation models, debugged the final simulation program, etc. I propose to formalize this (tentative) knowledge in the form of a regression model (this model must be tested later on to check its validity; see below). This regression model specifies which *inputs* and which *interactions* among these inputs seem important. These inputs are not only simulation parameters (e.g., service and arrival rates) and variables (say, number of servers) but also 'behavioral relationships' (like priority rules). 'Interaction' means that the effect of an input depends on the values of another input. So a tentative regression metamodel may be

$$y = \beta_o + \sum_{j=1}^{k} \beta_j x_j + \sum_{j=1}^{k} \sum_{g=j}^{k} \beta_{jg} x_j x_g + e\,, \qquad (1)$$

where β_o is the overall simulation response; β_j is the main or first-order effect of simulation input j; β_{jg} is the two-factor interaction between the inputs j and g ($g \neq j$); β_{jj} is the quadratic effect of input j (curvature); e denotes fitting or approximation error; y is 'the' response of a simulation run. Actually, a simulation run yields a time series of output values. That time series, however, can be summarized by one or more characteristics such as its average, its maximum, or its value at the end of the run. Moreover, the simulation model has several responses that are of interest; e.g., average waiting time of customers and utilization degrees of servers. I propose to apply eq. (1) to each response type individually. More sophisticated multivariate regression analysis is not warranted; see Kleijnen and Van Groenendaal (1992, p. 164).

To estimate $\boldsymbol{\beta} = (\beta_o, \beta_1, \dots, \beta_k, \dots, \beta_{12}, \dots, \beta_{kk})'$ (the effects of the simulation inputs), we fit a curve to the simulation data (**X**, **y**) where **X** denotes the nxQ matrix of inputs and y denotes the vector of n simulation outputs; in this case $Q = 1 + k + k(k-1)/2 + k = 1 + k + k(k+1)/2$. The classic fitting criterion is *Least Squares*. This criterion yields the estimator

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'y, \tag{2}$$

where a necessary but not sufficient condition for **X** is $n \geq Q$; see the next section.

Once the regression model is calibrated (its $\boldsymbol{\beta}$ is estimated), the *metamodel's validity* must be tested. For deterministic simulation models I propose cross validation: delete one of the n input combinations and its output (drop x'_i, y_i), reestimate from the remaining simulation data

$(\hat{\boldsymbol{\beta}}_{-i} = (X'_{-i}X_{-i})^{-1}X'_{-i}y_{-i})$;

predict the deleted simulation response y_i through the reestimated regression model

$(\hat{y}_i = x'_i\hat{\boldsymbol{\beta}}_{-i})$;

'eyeball' the relative prediction errors $\hat{y}_i/y_i$: are these errors acceptable to the user? For random simulations I prefer a lack-of-fit F-test (developed by Rao), unless the simulation-responses are not normally distributed (then cross validation is better). See Kleijnen and Van Groenendaal (1992, pp. 154-158).

The magnitudes of the estimated input effects $\hat{\boldsymbol{\beta}}$ quantify the importance of the corresponding inputs. Applications are numerous in deterministic and in random simulation. One case study concerns a set of deterministic ecological models (non-linear difference equations) that represent the effects that different gasses have on the global temperature: 'greenhouse' effect. These models require sensitivity analysis to support the Dutch government's decision making. Details are given in Kleijnen, van Ham and Rotmans (1992).

Optimization of the simulated system can be tried through *Response Surface Methodology* (RSM). In the first experiment a small area is explored, and RSM uses a first-order regression metamodel (see eq. (1) with the double summation term removed). The steepest ascent path (a search direction perpendicular to the regression plane) is followed to determine the next local experiment. After a number of local experiments, the 'hill top' is reached, and a (hyper)plane does not fit that (curved) top. Then a more extensive experiment is executed, which is analyzed through a second-order model (see eq. (1)). An application is provided by a decision support system (DSS) for production planning, developed for a Dutch company. To evaluate this DSS, a discrete-event simulation model is built. The DSS has 15 controllable inputs that are to be optimized. The effects of these 15 inputs are investigated, using a sequence of experiments (also see next section). Originally, 34 simulation response variables were distinguished. These 34 variables, however, can be reduced to one criterion variable, namely productive machine hours, that is to be maximized, and one commercial variable measuring lead times, that must satisfy a certain constraint. See Kleijnen and Van Groenendaal (pp. 181-185).

In all experiments, analysts use models such as eq. (1), explicitly or implicitly. For example, if they change one input at a time, then (implicitly) they assume that all interactions are zero. Of course it is better to make the regression model explicit and to find a design that fits that model, as I shall show next.

3. Experimental Design

For pedagogical reasons I first discuss classic designs and then discuss new designs for the initial screening phase of simulation experiments.

The nxk design matrix $\mathbf{D} = (d_{ij})$ specifies the n combinations of the k inputs that are to be simulated. (In multi-stage experimentation such as RSM, this set of combinations is followed by a next set.) Classic experimental design theory gives designs that are both 'efficient' and 'effective'. *Efficiency* means that n, the number of input combinations or simulation runs, is 'small'. Below eq. (2) we saw the condition $n \geq Q$ (Q denotes the number of effects in the regression metamodel). For example, k + 1 runs suffice if the k inputs are assumed to have first-order effects only. A popular but inferior design implies that the analyst observes the base situation, and then changes one input at a time. For example, for three inputs this design is: base run (-, -, -) where - means that the input is at its base value (this row vector is row 1 of D), (+, - ,-) as input 1 is increased from its base value to its maximum value within the area of experimentation, (-, +, -) as input 2 is increased, and (-, - , +) for input 3. Note that β_o corresponds with a vector of n ones in **X** of eq. (2), so **X** is 4x4 in this example with three first-order effects and one overall mean.

Now consider the *fractional factorial* 2^{3-1} design: its four columns are (-, +, -, +), (-, -, +, +), (+, -, -, +) (these columns specify D). It is easy to check that the corresponding **X** is orthogonal. Hence eq. (2) reduces to the scalar expression

$$\hat{\beta}_{j'} = \sum_{i=1}^{n} x_{ij'}\, y_i/n \quad (j' = 0,1,\dots,k). \tag{3}$$

How can we choose between these two designs? Classic statistical theory assumes that the fitting errors e are *white noise*: e is normally and independently distributed with zero mean and constant variance (say) σ^2. Then the variance-covariance matrix for the estimated input effects is

$$cov(\hat{\beta}) = \sigma^2 (X' X)^{-1} .$$

It can be proved that an orthogonal matrix **X** is *'optimal'*. Actually, there are several optimality criteria; see Kleijnen (1987, p. 335). Anyhow, an orthogonal **X** minimizes the variances of the estimated effects (the elements on the main diagonal of eq.(4)). There are straightforward procedures for deriving 'good' design matrices in case n equals 2^{k-p} with p=0,1...; for other n values there are tables and software; see Kleijnen and Van Groenendaal (1992, pp. 167-186).

Classic designs are also *effective*: they permit the estimation of *interactions*. If the regression metamodel includes two-factor interactions, then Q (the number of effects) increases to $1 + k + k(k-1)/2$. For example, if $k = 5$ then a 2^{5-1} design specifies 16 simulation runs to estimate 16 effects. If the inputs are quantitative, then a second-order regression model includes k quadratic effects. In such a model, n must further increase and more than two levels per input must be simulated. See Kleijnen and Van Groenendaal (1992).

A case study concerns a Flexible Manufacturing System (FMS). The four inputs of the deterministic simulation consists of the 'machine mix', i.e., the number of machines of type j with j = 1,...,4. Intuitively selected combinations of these inputs give inferior results when compared with a classic design. The simulation data are analyzed through two different regression metamodels. These models are validated: a regression model with only two inputs but including their interaction, gives valid predictions and sound explanations. See Kleijnen and Van Groenendaal (1992, pp. 162-164).

For didactic reasons I present *screening* designs after the classic experimental designs. In practice, most simulation models have many inputs; of course the analysts assume that only a few inputs are really important (parsimony). So in the beginning of a simulation project it is necessary to search for the few really important inputs among the many conceivably important inputs.

One approach is *group screening*, introduced in the early 1960s by several authors. This technique aggregates the many individual inputs into a few groups. Some simulation applications can be found in Kleijnen (1987, p. 327); these applications are queuing simulations. Recently Bettonvil and Kleijnen (1992) further developed group screening into *sequential bifurcation*, which is a very efficient technique that accounts for white noise and interactions. They applied this technique to screen the greenhouse model (discussed above) with nearly 300 inputs!

More efficient but complicated approaches do not treat the simulation model as a black box, but use analytical differential analysis: *perturbation analysis* and *score function*; see Kleijnen and Van Groenendaal (1992, p. 181).

Experimental design theory concentrates on a single response variable (denoted by y in this paper). In practice we can use the resulting designs to specify simulation runs; next we observe *several* responses per input combination, and analyze these results through regression analysis, as explained in the preceding section. See Kleijnen and Van Groenendaal (1992).

4. Conclusions

Experimental design and regression analysis are statistical techniques that are gaining popularity in simulation. These techniques are used for sensitivity analysis and optimization of simulation models. A number of case studies have been published. The techniques need certain adaptations to account for the peculiarities of simulation. Their application to simulation is straightforward.

References

Bettonvil, B. and J.P.C. Kleijnen 1992. *Identifying the Important Factors in Simulation Models with Many Factors*. Tilburg: Tilburg University.

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

Kleijnen, J.P.C., G. van Ham and J. Rotmans, 1992. Techniques for sensitivity analysis of simulation models: a case study of the CO_2 greenhouse effect. *Simulation* 58, no. 5: 410-417.

Kleijnen, J.P.C. and W. van Groenendaal 1992. *Simulation: a Statistical Perspective*. Chichester (U.K.): Wiley.