APPLIED MULTIRESPONSE METAMODELING FOR QUEUING NETWORK SIMULATION EXPERIMENTS: PROBLEMS AND PERSPECTIVES

Panajotis Katsaros Eleftherios Angelis Constantine Lazos

Department of Informatics
Aristotle University of Thessaloniki
54006 Thessaloniki
Greece
{katsaros, lef, clazos}@csd.auth.gr

ABSTRACT

A complete performance evaluation study of a simulated system should consider possible alternatives and response predictions to potential parameter changes. Simulation sensitivity analysis and metamodeling constitute an efficient approach for this kind of problems. However, this approach is usually despised, mainly because, a sophisticated methodological treatment is required. Such a methodology should take into account, peculiarities, inherent to queuing network models, as for example, multiple responses, large number of model parameters, many qualitative parameters etc. This work aims to illustrate the combined use of the proper statistical techniques to cope with this sort of problems and to show the need for a sound methodological framework that will bring this approach closer to the queuing network simulation practice.

KEYWORDS: Queuing networks, Simulation, Metamodeling

1 INTRODUCTION

Metamodeling is an efficient approach for studying the characteristics of a simulation model representing a system. It provides the means for model validation and verification and can be used for the prediction of system performance and the optimum operating conditions. However, metamodeling is usually done in an ad hoc way. Only recently we have seen (Kleijnen & Sargent, 2000) an interest in specifying a general methodology for developing metamodels.

A sound methodological framework with detailed process descriptions for applying metamodeling in queuing network simulations is required, since particular problems arise in these studies. More precisely, some of the common problems encountered in simulation studies of queuing networks are:

- multiple performance measures, such as, mean response times, throughputs, resource utilizations etc.
- qualitative input parameters, as for example, queuing disciplines and priority routings.
- nonlinear relationships between input and output variables, as for example, the observed dramatic variations of performance measures under small or moderate changes of parameters (Cheng & Kleijnen, 1999)

In this paper, a pilot study is used, in order to present a practical comprehensive approach, based on appropriate statistical techniques and to identify, where and how a methodological framework should support a safe statistical inference procedure. Certainly, the system model used for the present study does not reflect the whole range of the aforementioned problems. More realistic and complex models have to be analyzed, if a detailed set of guidelines is to be stated. The work of Kleijnen & Sargent (2000) was a valuable high-level base for our study, which aims to:

- interpret the effects of the model’s input parameters on the performance measures of interest
- predict the model’s performance in hypothetical situations and analyze scenarios regarding the architecture and the design of the system
- find ways to determine the optimal structure of the system, taking into account the inherent limitations of the available resources

The paper is organized as follows: Section 2 describes the queuing network model used for our study. In Section 3 we discuss the procedures employed for the simulation’s output data collection. Section 4 outlines the results of the conducted metamodeling study. Finally, in Section 5 the conclusions of our work are presented along with directions for future research work.

2 DESCRIPTION OF THE EXPERIMENT

A simple (non analytically solvable) queuing network was used as the performance model of interest. The model represents a central computing process by a closed workload with simultaneous resource possession (memory) and probabilistic job routing characteristics (Figure 1).

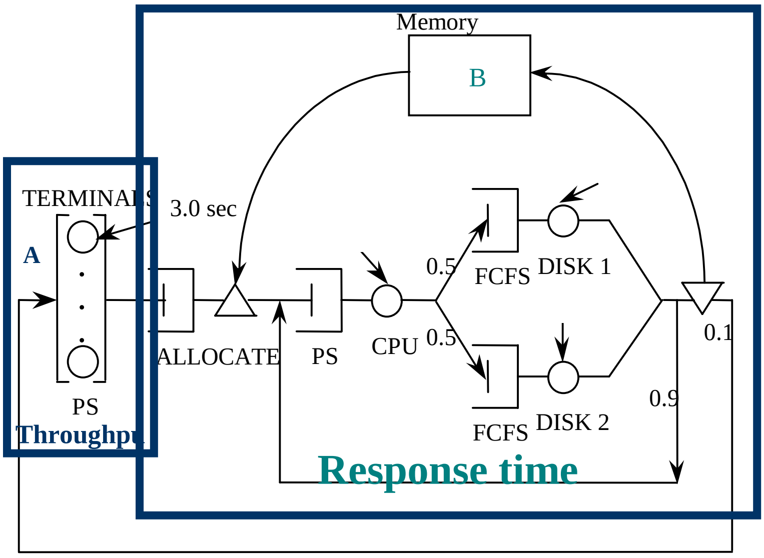

Figure 1. The performance model of a central computing process

The input variables, which influence the performance measures of interest, will be referred from now on as *independent variables* or *factors.* These are:

A: the number of terminals being served by the system
B: the number of memory partitions
C: the CPU speed
D: the speed of each of the disk(s) used for I/O
E: the number of disks (equally utilized)

The performance measures (dependent variables) that best characterize the system’s responsiveness for our study are:

- the mean system’s throughput (THRPUT), defined as the mean number of jobs passing through the terminals in the unit of time and,
- the mean system’s response time (RESPONSE).

The performed experiment, is based on the assumption, that each of the previously mentioned factors can take three possible values (*levels*), considered as representing the available resource alternatives. The workload (terminals) was varied over a range of values from a light to a heavily loaded case. The chosen levels of all the factors are shown in Table 1.

Table 1. The factor settings at which the simulation runs were performed

Factor	Levels	Units
Number of terminals (A)	10, 25, 40	terminals
Number of memory partitions (B)	2, 4, 6	partitions
CPU speed (C)	0.009, 0.012, 0.015	secs per job
Disk speed (D)	0.023, 0.028, 0.033	secs per job
Number of disks (E)	1, 2, 3	disks

Since the objective of the experiment was to investigate the main effects and the second order interactions of the five factors (A-E) on the dependent variables THRPUT and RESPONSE, a 3^{5-1} factorial experimental design (81 simulation runs) was chosen from Colbourn & Dinitz, (1996, p.349). Note that a complete factorial experiment would require 243 runs to cover all possible combinations. The selected experimental design is balanced in the sense that the replication numbers of all the factor levels are equal and this is also true for the concurrences of the levels of different factors.

The simulation runs, took place by using the HIT tool. HIT (Beilner, 1989; Beilner et al 1988; Beilner & Stewing 1987) is a performance analysis tool, which employs hierarchical modeling techniques for the structured specification and the quantitative (analytical or simulation based) evaluation of computing system models.

3 COLLECTION OF THE SIMULATION OUTPUT DATA

Data credibility is a critical issue for any metamodeling study. In queuing network simulations, this problem takes the form of correctly starting and stopping the runs, to achieve the required statistical accuracy. Since our aim is the steady-state behavior of the performance measures of interest, a sufficiently large set of observations has to be used for each simulation run. Moreover, an appropriate estimation procedure has to either remove or take into consideration the actual correlations among the generated observations. HIT is one of the few available simulation packages that offer such a statistical preprocessing of the produced simulation result. More precisely, the *method based on autoregressive representation* (Litzba et al, 1989) has been used, for transforming the originally correlated observations into a sequence of independent and identically distributed random variables. A thorough comparative study of other techniques that can be used for this purpose is given in (Pawlikowski, 1990).

The method's implementation is equipped with the *relative confidence interval precision* criterion, which has been used as the stopping rule for the conducted experiments. Let us denote by $\left(\overline{X}(n)-\Delta_x, \overline{X}(n)+\Delta_x\right)$, the generated (autoregressive method) interval, which contains the true parameter μ_x at a given confidence level (1-α), 0<α<1. If

$$\delta = \frac{\Delta_x}{\overline{X}(n)} \tag{1}$$

is the relative half width of the confidence interval, then the simulation experiment is stopped at the first checkpoint for which $\delta \leq \delta_{\max}$, where $\delta_{\max}$ is the required limit relative precision, at the $100(1\text{-}\alpha)\%$ confidence level. For the purposes of our study we used $\delta_{\max} = 0.01$ and $\alpha = 0.05$.

4 METAMODELING ANALYSIS

In order to achieve the objectives outlined in Section 1, certain exploratory and inferential statistical techniques were employed. More specifically:

- the interpretation of the factor effects on the performance measures was accomplished by the use of appropriate techniques such as

  - mean value plots, box plots and interaction plots
  - analysis of variance (ANOVA)
- prediction of the model’s performance was carried out by
  - scatterplots and curve fitting for exploring the relationship between the dependent variables
  - regression metamodels and response surfaces for both dependent variables
- identification of the optimal system configuration was carried out by
  - regression metamodeling for a chosen cost function combining both performance measures
  - techniques for finding local or global minimum cost values such as analytic solutions based on derivatives, the steepest descent procedure and a simulated annealing algorithm

The two plots of Figure 2, show the mean values of the variables RESPONSE and THRPUT respectively for each level of the factors, with respect to the overall mean (horizontal line). We can see that the CPU (C) and the disk speeds (D) have a linear and relatively small effect on the mean of the dependent variables, while the other factors effect, is generally nonlinear. The factor A (number of terminals) seems to have the largest effect.

The box plots presented in Figure 3, depict the effect of the number of terminals (A), on the whole distributions of the dependent variables. They indicate that, the increase of the number of terminals has a significant effect, not only on the location of the data, but also, on their dispersion. This tool is ideal for identifying skewness and asymmetry in the data, as well as, for detecting outliers (Heidelberger & Lewis, 1981).

The plots presented in Figure 4 illustrate the combined effects of factor A with B and E respectively, on the response time. It seems that, for the smallest number of memory partitions, the response time has a steepest increase, when the number of terminals increases.

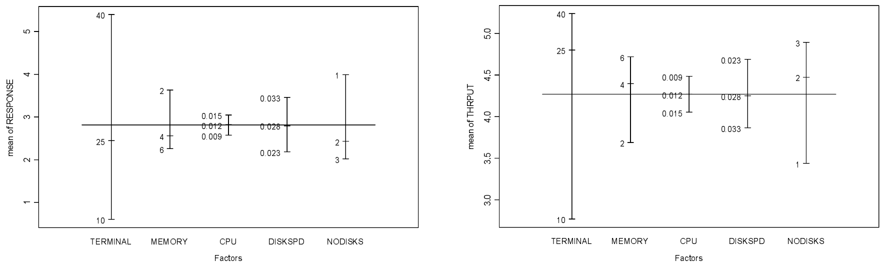

Figure 2. Mean value plots

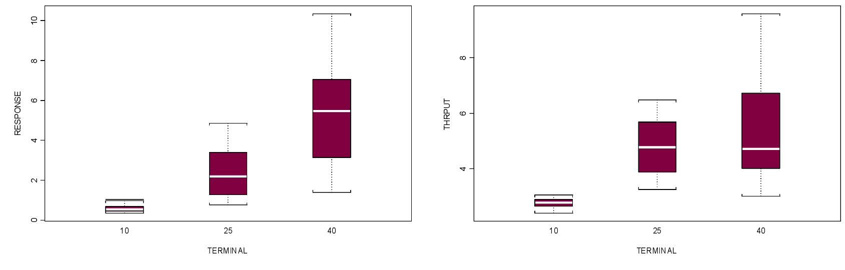

Figure 3. Box plots

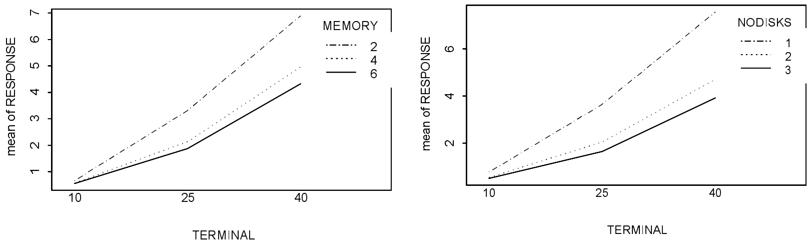

Figure 4. Interaction plots

Univariate Analysis of Variance (ANOVA) was conducted for both dependent variables in order to examine the main effects of the five factors and their 2-way interactions. RESPONSE and THRPUT were found to be significantly affected by all the factors A-E and the interactions AB, AD, AE, BE.

In the scatterplot of Figure 5 (left), we identified three clearly distinguishable sets of points, corresponding to the levels of factor A. Based on this grouping, we tried to fit a number of regression curves describing the relation between the two dependent variables (Figure 6). Thus, in the case of 25 terminals the best fitting metamodel was found to be:

$$RESPONSE \;=\; \text{-}3.41 + (26.88/THRPUT) \text{ with } r^2 = 0.997 \tag{2}$$

The next step in our analysis was the generation of response surfaces using least squares regression. For this purpose, the factors A, B, C, D and E were considered as quantitative variables and their levels as numerical values. For the construction of the appropriate model we used certain regression procedures (backward and stepwise regression). We tried a large number of models, considering not only the original variables, but also their squared values and their pair-wise products. We concluded in two exponential models, with r^2=0.985 and 0.981 respectively:

$$\begin{aligned} RESPONSE = \exp(&-2.031 + 0.146A - 0.125B + 46.544D - 0.800E - 0.001019A^2 + 0.02727B^2 \\ &+ 0.156E^2\text{-}0.003306A\cdot B - 0.005128A\cdot E - 0.06563B\cdot E + 18.713C\cdot E) \end{aligned} \tag{3}$$

$$\begin{aligned} THRPUT = \exp(&0.518 + 0.07458A - 14.065D + 0.193E - 0.001023A^2 - 0.008498B^2 - 0.08235E^2 \\ &+ 0.002543A\cdot B - 0.511A\cdot C - 0.706A\cdot D + 0.006156A\cdot E \\ &+ 0.02692B\cdot E + 499.882C\cdot D - 7.296C\cdot E + 4.148D\cdot E) \end{aligned} \tag{4}$$

Figure 5 (right) shows predictions, based on the aforementioned metamodels for different numbers of terminals (A=10, 15, 20, 25, 30, 35, 40) and for all the level combinations of the other factors.

A sample response surface for THRPUT, based on the regression model (4) and the corresponding contour plot, is presented in Figure 7. These plots are used for examining certain questions, regarding the system behavior when some of the factors are kept fixed, while others variate within predefined limits.

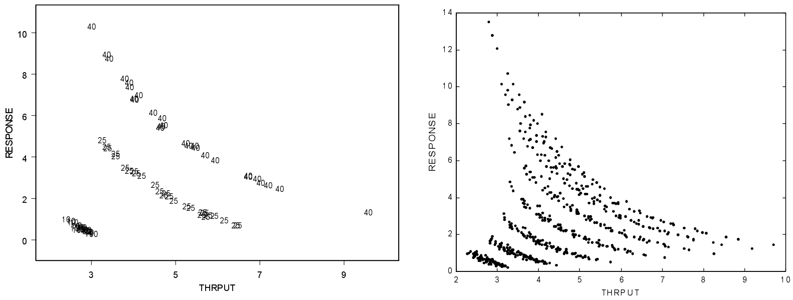

Figure 5. Actual and estimated values of the dependent variables

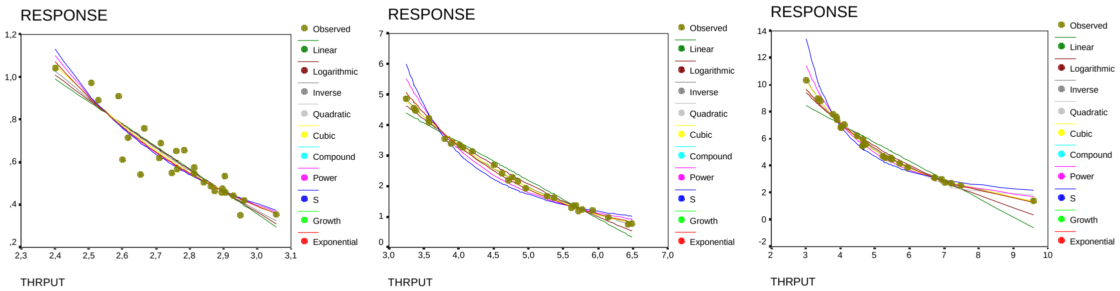

Figure 6. Curves describing the relationship between the dependent variables

Finally, a cost function, defined as the ratio of RESPONSE over THRPUT, was decided that best describes the system responsiveness and served as the criterion for the conducted optimization studies. Stepwise regression analysis resulted in an exponential regression model. The obtained function was then used to find the optimum system configuration in different hypothetical situations. The techniques used, were the classical calculus-based methods (analytical solution, steepest descent) and also a simulated annealing algorithm (Bohachevsky et al, 1986). In all cases, the expected results were obtained and this confirms the validity of the used metamodel relative to the system model.

Statistical analysis was conducted by using the SPSS, Splus and MATLAB software tools.

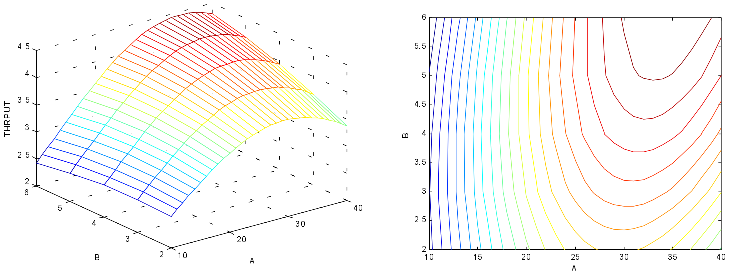

Figure 7. Example of surface and contour plots for fixed values of factors C, D and E

5 CONCLUSION

In this paper we present the combined use of appropriate techniques for a complete metamodeling study. The objective was to describe the performance of a queuing network system, by models obtained from statistical analysis on the results of simulation runs.

We first introduced the specific aims of our study. Then the input and output variables of interest were identified and the values representing the available resource alternatives were determined. We employed an experimental design and the simulation runs were executed according to it. The results were obtained in a pre-specified accuracy level, achieved by the use of an appropriate technique (autoregressive).

Various statistical techniques were used, in order to explore the effect of the input variables on the system's responsiveness. Prediction and optimization of the model's performance, was based on a set of well-fitted regression models.

We believe that the present study, constitutes a base for a future research effort, in adapting the techniques presented, for use in more realistic queuing network models.

REFERENCES

Beilner, H. (1989). Structured Modeling – Heterogeneous Modeling, *Proc. of the 1989 European Simulation Multiconference*, Rome

Beilner, H., Mater, J., Weiβenberg, N. (1988). Towards a performance modeling environment: News on HIT, *Proc. of the 4th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation*, Palma de Mallorca, Plenum Publishing Corporation

Beilner, H., & Stewing, F. J. (1987). Concepts and techniques of the performance modeling tool HIT, *Proc. of the European Simulation Multiconference, ESM'87*, Vienna, Austria

Bohachevsky, I. O., Johnson, M. E., Stein, M. L. (1986). Generalized simulated annealing for function optimization, *Technometrics*, 28/3, 209-217

Cheng, R. C. H., & Kleijnen J. P. C. (1999). Improved design of queuing simulation experiments with highly heteroscedastic responses, *Operations Research*, 47/5, 762-777

Colbourn, J. C., & Dinitz J. H., (1996). *The CRC handbook of combinatorial designs*, CRC Press, Inc.

Heidelberger, P. & Lewis, P. A. W. (1981). Regression – Adjusted estimates for regenerative simulations with graphics, *Communications of the ACM*, 24/4, 260-273

Kleijnen, J. P. C., & Sargent, R. G. (2000). A methodology for fitting and validating metamodels in simulation, *European Journal of Operational Research*, 120, 14-29

Litzba, D., Sczittnick, M., Stewing, F. J. (1989). Yet another simulation output analysis algorithm: the autoregressive, online-update evaluation technique of the modelling tool HIT, *Proc. of the 3rd European Simulation Congress*, Edinburgh

Pawlikowski, K. (1990). Steady-state simulation of queuing processes: A survey of problems and solutions, *ACM Computing Surveys*, 22/2, 123-170