Challenges at the stage of simulation experiment planning

O.M. Demidenko, N.A. Aksionova, A.V. Varuyeu
F. Skorina Gomel State University, Sovetskaya str, 104, Gomel, Belarus,
demidenko@gsu.by, nataliaksen@gmail.com, varuyeu@gmail.com

Abstract*. The article describes an approach to optimizing the process of modeling the computational process. The main emphasis is placed on the organization of simulation experiments and their reorganization depending on the objectives of the study. An algorithm for choosing the number of sample experiments to satisfy the attainability criterion for all components of the accuracy vector is formulated. An approach to organizing the search for the culprit of the simulation model error and the place of its occurrence in the model algorithm is outlined. It is concluded that there are no restrictions on the type of dependencies between the parameters of the simulation model and the state of the external environment of a complex system.*

I. Introduction

Once the experiment goals are set, the decision on use of the computer-aided simulation is made and the simulation automation system is defined, pre-planning of a prospective experiment is feasible [1-7]. The investigator shall have a detailed experiment plan for targeted and efficient data acquisition. The expected time and expense limits shall be adjusted to the investigator's available resources. The more expensive and complex the experiment is the more attention shall be paid to the named stage. On frequent occasions the restrictions in experiment resources are so severe, that traditional statistical methods become unavailable.

II. "Automatic stop" rule

The rule is based on the method of confidence intervals. The method of confidence intervals includes determination of accuracy d_n of expectation function μ_n or b_n for the dispersion of σ_n^2 *n*-th component of response vector Y and of significance level α, providing for $\mu_n\sigma_n^2$ being well within intervals $(Y_n + d_n), (D_n \pm b_n)$ with probability of $(1 - \alpha)$. Y_n and D_n here are the mean value and the dispersion calculated by the sample with size N and are the assessments of μ_n and σ_n^2 respectively. In the course of testing and investigation of simulation properties, the investigator determines accuracy vectors $(d_1, \dots, d_n, \dots, d_L)$ or $(b_1, \dots, b_n, \dots, b_L)$ of the representation of the component of response vector Y. Implementation of the "automation stop" rule is an iterative procedure, the essence of which is described below. Significance level α and accuracy vectors of representation of the component of response vector (b_n) and (d_n) are known prior to commencement of the experiment series. The number of initial experiments N_1 needed to generate samples of the component values of model responses $\{Y_{nk}\}$, $n = \overline{1, L}$; $k = \overline{1, N_1}$ is established based on a-priori data, such as the experience generated during investigation of the simulation properites. If the investigator supposes that amount of experiments of the investigation of simulation properties is too high, he/she establishes $N_1 = 5$ as the initial value. The subsequent algorithm of selection of the number of experiments N consists of the following steps.

Step 1. Mean values $\bar{Y}_n$ and dispersions D_n are determined based on sample $\{Y_{nk}\}$.

Step 2. For subsequent number n the vector components of model response Y determine achieved assessment accuracy $\bar{Y}_n$ and D_n at N_1executed experiments. There can be a number of different cases. If the sample is of small size $(N_1 \leq 30)$, *t*-statistics with Student's distribution $d_{1n} = t_{kp}\sqrt{D_n}/(N_1 - 1)$, where t_{cr} is the critical value of *t*-statistics determined by the table of Student's distribution [2] at N1 – 1 degree of freedom and the given significance level α, is used for calculation of the confidence interval. If the sample size is big $(N_1 > 30)$, two-sided statistics with standardized normal distribution is used for calculation of the confidence interval.

$$d_{1n} = z_{\alpha/2}\sqrt{D_n/N_1}, \quad (1)$$

where $z_{\alpha/2}$ is the value of the standardized normal distribution, which is determined using the table [2] at the given significance level $\alpha/2$.

If normality Y_n cannot be assumed, but the value is very high, Chebyshev inequality is used:

$$p\{|\bar{Y}_n - \mu_n| \geq h\sigma_n/\sqrt{N_1}\} \leq 1/h^2, \quad (2)$$

where h is some pre-determined constant, i.e. the value of root-mean-square deviations meeting the investigator's requirements. The confidence interval then may be calculated with sufficient accuracy using the formula

$$d_{1n} = \sqrt{D_n}/N_1(1-\alpha), \quad (3)$$

when assessing the dispersion at search for the assessment of D_n with accuracy $(1-\alpha)$, the below inequality shall be met:

$$p\{(1-b_n)\sigma_n^2 \leq D_n \leq (1+b_n)\sigma_n^2\} = 1-\alpha. \quad (4)$$

It is feasible to solve it at high number of experiments N_1. The chi-square statistics is used in this case. Since N_1 is quite high, the statistics is approximated based on the normal distribution to generate the following formula for determination of the achieved assessment accuracy σ_n^2 in N_1 experiments:

$$b_{1n} = z_{\alpha/2}\sqrt{2/(N_1-1)}, \quad (5)$$

where $z_{\alpha/2}$ is the value of the standardized normal distribution, which is determined using the table at the given significance level $\alpha/2$.

Step 3. The obtained accuracy d_{1n} or b_{n} is compared. If inequality $d_{1n} \leq d_n$ or $b_{1n} \leq b_n$ is true, the required assessment accuracy is reported as achieved on the *n*-th component of the response vector of Y model in N_1 experiments. If the inequality is not true, then proceed to step 4. When the inequality is true, proceed to step 5.

Step 4. Another model experiment is executed: N_1 is increased by one ($N_1 \equiv N_1 + 1$) and step 1 is proceeded.

Step 5. All components of the response vector are verified to meet the assessment accuracy μ_n or σ_n^2. If the accuracy is verified for all components of the accuracy vector $(d_1, \dots, d_n, \dots, d_L)$ or $(b_1, \dots, b_n, \dots, b_L)$, then the experiments are completed. Otherwise, the component number of the model response vector is changed and step 2 is repeated.

III. Selection of measurement intervals for simulation statistics

When developing a plan for analyzing the simulation results, the investigator shall be able to conduct one or more checks of accuracy and acceptability of the simulation results. In some cases, the balance equation can be used to determine the sources of errors. The essence of the name method is considered on the following example.

Let variables A, B, X, Y be measured in the model experiment, which are interconnected by the equation of conservation between the pairs of variables $A \cdot B = X \cdot Y$.

It has been established that due to inaccuracy of the description of the real process, one of these variables (but it is not known which one) is the cause of the systematic error. To correct the simulation experiment, not only the culprit of the error needs to be identified, but also the place of error occurrence in the model algorithm. Let A not change during the simulation, the remaining variables change as follows: $m \cdot X; n \cdot Y; m \cdot n \cdot B$. The balance equation under these conditions is as follows: $A \cdot m \cdot n \cdot B = m \cdot X \cdot n \cdot Y$.

The following rule applies: if one of the variables in the balance equation can be presented as a sum $(A + f(A))$, where $f(A)$ is a systematic error, then this variable can be found by considering one by one cases with a fixed value of each variable. A variable, the relative error of change of which does not change with its fixed value, is the cause of the systematic error. The exception is the case when the variable has the form $A + k \cdot A$ (k is a constant value). In this case, the error cannot be detected using this method.

Often, a preliminary experiment is used to find the simulation error. Based on the results of this experiment, the investigator plots the dependence $Y = f(X)$. The coordinate system and the function itself are selected in such a way that the plot is linear or at least does not have a large curvature near the origin for subsequent extrapolation. Linear and semi-logarithmic scales are commonly used to plot a straight line. In the case when there is a sharp deviation from the linear form in the range of small values, the data cannot be extrapolated, so the investigator shall monitor this situation.

Let the investigator know that one of the linear plots of the model response has a systematic error. It is necessary to determine which of the plots is correct, and to assess this error. For example, each set of simulation results is characterized with the same exponent, but the plots do not match. However, it is known that the dependence shall pass through the origin of coordinates in the range of small values. Then the plot that does not satisfy the condition of passing through the origin has a systematic error. The mean value of the coordinate at $X = 0$ is an estimate of this error.

IV. Conclusion

In the article, the authors considered in detail such aspects of simulation experiment planning as determination of the required sample size; determination of intervals of model parameter changes; planning of the search for error sources in the simulation experiment.

The simulation experiment itself is a very resource-intensive research method. Therefore, the final results of the analysis can be obtained by considering the sequence of different aspects of functioning of component groups of complex systems. At the same time, the advantages include: ability to describe behavior of the components of a complex system at a high level of detalization; absence of restrictions on the type of dependencies between the parameters of the simulation model and the state of the external environment of a complex system; the possibility of complex studies of the dynamics of interaction between the components of a complex system in time and space of the parameters of a complex system.

References

[1] *Demidenko O.M., Maksimey I.V.* Simulation modeling of effects in computers / O.M. Demidenko. – Minsk: Belarusian science, 2000. – 230 p. – ISBN 985-08-0439-4.

[2] *Demidenko O.M., Maksimey I.V.* Design modeling of the computational process in the search for computational applications / O.M. Demidenko. – Minsk: Belarusian Science, 2001. – 252 p. – ISBN 985-08-0460-2.

[3] *Demidenko O.M.* The technology of monitoring and adaptation of the computing process is selected for the local area network / O.M. Demidenko. – Minsk: Belarusian Science, 2002. – 193 p. – ISBN 985-08-0514-5.

[4] *Demidenko O.M.* "Techniques of Adapting a Calculating Process to Operating Load of a Local Area Network," 2022 International Conference on Information, Control, and Communication Technologies (ICCT), 2022, pp. 1-6, doi: 10.1109/ICCT56057.2022.9976851.

[5] *Барабанова Е.А., Мальцева Н.С.* Имитационное моделирование коммутационных систем//Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2009. № 1. С. 146-150.