Int j simul model 20 (2021) 3, 477-488
Original scientific paper

A ROBUST METHOD FOR IDENTIFYING THE BEST AND WORST SUBSETS IN STOCHASTIC SIMULATION

Choi, S. H.[*] & Kim, B. S.[**,#]

[*] Department of IT Convergence and Application Engineering, Pukyong National University, Busan, 48513, South Korea

[**] School of Software Convergence, Myongji University, Seoul, 03674, South Korea

E-Mail: shchoi@pknu.ac.kr, kimbs@mju.ac.kr ([#] Corresponding author)

Abstract

Stochastic simulation is a useful tool for evaluating modern discrete-event dynamic systems; however, efficiency issues still exist in simulation experiments because many simulation replications are required to estimate the mean performance of a system configuration accurately. This paper proposes a robust method to identify the *m* best and *n* worst candidates from a finite set of configurations according to their mean performance in a high-noise environment. To select the best and worst subsets correctly within a limited simulation budget, the proposed method defines a metric to evaluate the statistical significance of each candidate's current simulation results. Then, it allocates small further replications iteratively based on the calculated metric so the results become significant evidence to verify the correct identification. Experimental results on benchmark problems demonstrate the superior efficiency of the proposed method compared to the existing methods in high-noise situations. The proposed method is beneficial to best-worst scaling problems, multiple-criteria decision-making problems, population-based search algorithms, etc. Furthermore, it allows practitioners to make final decisions by considering qualitative criteria neglected by simulation.

(Received in February 2021, accepted in June 2021. This paper was with the authors 2 weeks for 2 revisions.)

Key Words: Stochastic Simulation, Simulation Experiments, Best and Worst Subsets, Simulation Budget Allocation, Robustness

1. INTRODUCTION

Stochastic simulation is a useful tool for evaluating the performance of modern discrete-event dynamic systems, such as transportation [1], manufacturing [2], supply chain management [3], etc. However, efficiency issues can arise when using simulation to optimize a system because many simulation replications are required to estimate precisely the mean performance of each candidate (i.e., simulation input, e.g., system configuration and scenario setting) to identify the best one accurately. If the number of candidates is finite, then efficient allocation methods based on the ordinal optimization have been developed to select the best one correctly within a limited number of simulation replications [4]. To maximize the statistical evidence representing the correct selection of the best candidate, these methods allocate a few further replications sequentially and iteratively based on the statistical comparisons and inferences for the candidates' current simulation results.

For example, the two-stage indifference-zone method [5] allocates further replications to the candidates of which the estimated mean performance is statistically similar to the current best one to achieve a predefined lower bound guarantee of the probability of correct selection (i.e., $P\{\text{CS}\}$). On the other hand, Chick et al. [6] aim to minimize the upper bound of the expected opportunity cost (i.e., $E[\text{OC}]$) using decision-theory tools for allocation. The optimal computing budget allocation (OCBA) method [7] allocates further replications with its allocation rules derived to maximize the lower bound of $P\{\text{CS}\}$ asymptotically. Similar to OCBA, the uncertainty evaluation (UE) method [8] also aims to maximize $P\{\text{CS}\}$, but its allocation is based on the uncertainty of each design that evaluates the statistical significance of their current simulation results.

Most previous allocation methods have focused on finding only the best candidates, but practitioners in engineering may be interested in the worst one as well as the best one. For example, when analysing the stability of a production system, it is necessary to consider simultaneously the best and worst scenarios in which the performance is minimized. The best-worst scaling, a discrete choice model used in a variety of fields (e.g., psychology, food science, and healthcare) requires picking both the highest and lowest utility items from *k* candidates at the same time under conditions involving randomness and noise [9]. In addition, selecting the best and worst is beneficial to multiple-criteria decision-making problems by constructing the ideal and nadir points for each criterion [10]. To meet this need, Zhang et al. [11] and Choi and Choi [10] proposed efficient allocation methods based on OCBA and UE methods, respectively.

In practice, finding only the best and worst candidates may not be sufficient for practitioners because the simulation model is an abstraction of real systems; thus, it is necessary to identify the best and worst subsets, which is a generalization of the best and worst selection. The best and worst subsets consist of the top-*m* best candidates and the top-*n* worst candidates, respectively, so that practitioners can make final decisions by considering some conditions neglected by simulation (e.g., qualitative criteria or political feasibilities) as well as the simulation results [12]. This is particularly useful in solving best-worst scaling problems that require multiple performance measures [13]. In addition, identifying the best and worst subsets can improve the performance of population-based search algorithms (e.g., genetic algorithm) in stochastic environments because the correctly selected best and worst subsets help to increase the quality of the population with the selection schemes [14]. Despite such utility, little research exists on identifying the best and worst subsets. Xiao et al. [13] developed an OCBA-based method called OCBAmn by extending the previous method [11] for selecting the best and worst candidates, and Zhang et al. [15] generalized OCBAmn to split *k* candidates into ranked subsets.

Recently, with increasing complexity of modern industrial systems, stochastic simulation models have also become complex. As a result, the cost per simulation replication is expensive, and the simulation output has high variance. In addition, since the emergence of the digital twins, which constantly control real systems based on simulation models synchronized with the systems, the need to identify the best and worst subsets more efficiently is increasing. Thus, we propose a robust method for identifying the best and worst subsets in stochastic simulation. Unlike OCBAmn, the proposed method is based on the UE method. Compared to the OCBA method, UE is relatively heuristic, but it is practical and has high robustness to noise. To maximize $P\{\mathrm{CS}\}$ under a limited simulation budget, the proposed method calculates the uncertainty of each candidate to evaluate their current simulation results using statistical hypothesis testing and *p*-values, and then sequentially and iteratively allocates a few further replications based on the uncertainty. Although OCBAmn considers only the sample mean and sample variance of the current simulation results in the allocation, the proposed method further utilizes the number of allocated replications so far. Thus, the proposed method can improve efficiency by minimizing wasted budget due to inaccurate sample mean in a high-noise situation, which is demonstrated in various experimental results on benchmark problems.

The remainder of this paper is organized as follows: Section 2 defines the problem, and Section 3 proposes the robust method for identifying the best and worst subsets. Section 4 exhibits the experimental results, and a conclusion is given in Section 5.

2. PROBLEM DEFINITION

Let x_i be a simulation input and Θ be a finite set of candidates consisting of *k* inputs (i.e., $\Theta = \{x_1, x_2, \ldots, x_k\}$). Suppose the simulation output sample $\hat{y}_i$ of x_i in a simulation replication follows an independent and identically distributed normal distribution with an unknown mean μ_i and variance σ_i^2 (i.e., $\hat{y}_i \sim N(\mu_i, \sigma_i^2)$), which is a common assumption in the literature [4, 7].

This is reasonable in stochastic simulation because $\hat{y}_i$ is typically obtained as an average value or batch mean, thus the central limit theorem holds. In addition, we assume that no prior knowledge of μ_i and σ_i^2 is given before conducting a simulation to meet practical situations.

To simplify the problem, no candidates have the same mean performance (i.e., $\mu_i \neq \mu_j$, $i, j \in \{1, \dots, k\}$, and $i \neq j$), and the candidate with the smallest μ_i is the best. Then, the best subset S_B and worst subset S_W that consist of the top-m best candidates and the top-n worst candidates, respectively, can be defined as follows:

$$S_B = \left\{ x_i \left| \max_{x_i \in S_B} \mu_i < \min_{x_j \notin S_B} \mu_j, i, j \in \{1,\dots,k\}, \text{and } |S_B| = m \right. \right\} \tag{1}$$

$$S_W = \left\{ x_i \left| \max_{x_i \notin S_W} \mu_i < \min_{x_j \in S_W} \mu_j, i, j \in \{1,\dots,k\}, \text{and } |S_W| = n \right. \right\} \tag{2}$$

This paper does not consider the rank order within the subset. However, the unknown mean performance μ_i can only be estimated with the sample mean $\bar{\mu}_i$ of the collected N_i simulation output samples (i.e., $\bar{\mu}_i = 1/N_i \sum_{j=1}^{N_i} \hat{y}_{ij}$); thus, in practice, we inevitably identify the estimated best and worst subsets defined as below:

$$\bar{S}_B = \left\{ x_i \left| \max_{x_i \in \bar{S}_B} \bar{\mu}_i < \min_{x_j \notin \bar{S}_B} \bar{\mu}_j, i, j \in \{1,\dots,k\}, \text{and } |\bar{S}_B| = m \right. \right\} \tag{3}$$

$$\bar{S}_W = \left\{ x_i \left| \max_{x_i \notin \bar{S}_W} \bar{\mu}_i < \min_{x_j \in \bar{S}_W} \bar{\mu}_j, i, j \in \{1,\dots,k\}, \text{and } |\bar{S}_W| = n \right. \right\} \tag{4}$$

The selected $\bar{S}_B$ and $\bar{S}_W$ based on $\bar{\mu}_i$ may be different from S_B and S_W, respectively. If every $\bar{\mu}_i$ becomes precise with many simulation replications for each candidate, then $\bar{S}_B$ and $\bar{S}_W$ can be correct, but this brings about the simulation efficiency problem. The total number of replications T is typically limited in the form of computing resources or time; thus, to increase the efficiency, it is necessary to allocate T to each candidate selectively such that $\bar{S}_B$ and $\bar{S}_W$ become correct. When $P\{\text{CS}\}$ is used for the statistical evidence representing the correct identification of $\bar{S}_B$ and $\bar{S}_W$, this can be defined as the following optimization problem:

$$\arg\max_{N_1,\dots,N_k} P\left\{ \left(S_B = \bar{S}_B\right) \wedge \left(S_W = \bar{S}_W\right) \right\} \quad \text{s.t. } N_1 + N_2 + \cdots + N_k = T \text{ and } N_i \geq 0 \tag{5}$$

This problem implicitly assumes the simulation cost per replication across candidates is similar. As mentioned previously, to solve this problem, this paper proposes a robust allocation method to maximize $P\{\text{CS}\}$ under the constraint of T based on the UE method.

3. PROPOSED METHOD

In common with the existing methods, the proposed method divides a given budget T into small pieces and allocates them sequentially and iteratively to use T efficiently. Concretely, because no prior knowledge of simulation results for k candidates is given, initial n_0 replications are allocated to each candidate to obtain the minimum information. Then, based on the information, the proposed method allocates Δ further replications to k candidates depending on the uncertainty and updates their simulation results. After repeating allocations and updates until T is depleted (i.e., $(T - kn_0)/\Delta$ times), $\bar{S}_B$ and $\bar{S}_W$ are selected based on the current results and returned as the best and worst subsets (see Algorithm 1).

To maximize $P\{\text{CS}\}$ within T, Δ replications have to be allocated so that selected $\bar{S}_B$ and $\bar{S}_W$ are the same as S_B and S_W, respectively. According to Eqs. (1) and (2), for the correct selection of $\bar{S}_B$ and $\bar{S}_W$, the unknown mean performance μ_i of each candidate x_i ($i \in \{1, \dots, k\}$) has to meet the following relationships:

$$\left[\left(\underline{\mu_{i^1} < \mu_{i^{m+1}}}\right) \wedge \left(\underline{\mu_{i^1} < \mu_{i^{m+2}}}\right) \wedge \cdots \wedge \left(\underline{\mu_{i^1} < \mu_{i^k}}\right)\right] \wedge \cdots \wedge \left[\left(\underline{\mu_{i^m} < \mu_{i^{m+1}}}\right) \wedge \cdots \wedge \left(\underline{\mu_{i^m} < \mu_{i^k}}\right)\right]$$
$$\wedge \left[\left(\underline{\mu_{i^{m+1}} < \mu_{i^{k-n+1}}}\right) \wedge \cdots \wedge \left(\underline{\mu_{i^{m+1}} < \mu_{i^k}}\right)\right] \wedge \cdots \wedge \left[\left(\underline{\mu_{i^{k-n}} < \mu_{i^{k-n+1}}}\right) \wedge \cdots \wedge \left(\underline{\mu_{i^{k-n}} < \mu_{i^k}}\right)\right] \quad (6)$$

The superscript of the subscript in $\mu_{i^1}, \ldots, \mu_{i^k}$ represents the estimated ranking of x_i based on $\bar{\mu}_i$. In other words, the ranking of candidates belonging to $\bar{S}_B$ is from 1 to m (i.e., $\bar{S}_B = \{x_{i^1}, \ldots, x_{i^m}\}$), and that of candidates belonging to $\bar{S}_W$ is from $k-n+1$ to k (i.e., $\bar{S}_W = \{x_{i^{k-n+1}}, \ldots, x_{i^k}\}$). Simply put, Eq. (6) means μ_i of every x_i belonging to $\bar{S}_B$ should be smaller than that of the others that do not belong to $\bar{S}_B$, whereas μ_i of every x_i belonging to $\bar{S}_W$ should be greater than that of the others that do not belong to $\bar{S}_W$. In Eq. (6), duplicated relationships are excluded; thus, the total number of relationships is $(k-m)m + (k-m-n)n$.

If every relationship in Eq. (6) can be verified as true based on the current simulation results of k candidates, in other words, if the current results can be significant evidence to be statistically convinced that these relationships are true, no further simulation replications are required. On the other hand, if the results of candidates do not provide (relatively) significant evidence to verify any relationship in Eq. (6), then the results should be updated by assigning Δ to these candidates to maximize $P\{\mathrm{CS}\}$. The concept of uncertainty based on statistical hypothesis testing and p-value makes it possible to achieve such an allocation strategy [8].

Statistical hypothesis testing is a frequentist method to verify a relationship between unknown μ_i and μ_j using p-value calculated based on the observed simulation results, such as $\bar{\mu}_i$, $\bar{\mu}_j$, s_i^2, and s_j^2. Here, s_i^2 is the sample variance of the obtained simulation outputs for x_i (i.e., $s_i^2 = 1/(N_i - 1) \cdot \sum_{j=1}^{N_i} (\hat{y}_{ij} - \bar{\mu}_i)^2$). For example, the test for verifying $\mu_i < \mu_j$ is as follows:

$$H_0 : \mu_i > \mu_j, \quad \underline{H_A : \mu_i < \mu_j} \quad (7)$$

The relationship $\mu_i < \mu_j$ is set to the alternative hypothesis H_A, and its opposite is set to the null hypothesis H_0. Here, H_0 has no equal sign depending on the assumption that no candidates have the same mean performance. The p-value for the test is calculated as below:

$$p_{i,j} = F_v\left[\left(\bar{\mu}_i - \bar{\mu}_j\right) \Big/ \sqrt{s_i^2/N_i + s_j^2/N_j}\right] \quad (8)$$

The function F_v is the cumulative distribution function of the t-distribution with v degrees of freedom, where v is calculated as follows:

$$v = \left\lfloor \left(s_i^2/N_i + s_j^2/N_j\right)^2 \Big/ \left[s_i^4/\left(N_i^3 - N_i^2\right) + s_j^4/\left(N_j^3 - N_j^2\right)\right] \right\rfloor \quad (9)$$

If the variance of simulation output is known, then F_v in Eq. (8) can be replaced with the cumulative distribution function of the standard normal distribution.

The p-value $p_{i,j}$ is the probability of obtaining the observed simulation results (which are used for calculating $p_{i,j}$) or more extreme results (which are less likely to be obtained) under the assumption that H_0 is true. In the hypothesis testing, $p_{i,j}$ is used to indicate the statistical significance of the observed results to verify H_A. A value of $p_{i,j}$ close to zero means the observed results have a large statistical significance. That is, because obtaining the observed results is very rare under the assumption that H_0 is true (although they were actually obtained), the results can be significant evidence by which to reject H_0 and accept H_A. On the other hand, as $p_{i,j}$ increases, the significance of the results decreases. Namely, because the observed results are likely to be observed and more plausible under the assumption that H_0 is true, the results cannot be evidence by which to reject H_0. In this case, H_0 is accepted and H_A is rejected, but this does not mean H_A is false. It is an uncertain situation where H_A cannot be verified with the current observed results. Because $p_{i,j}$ is a probability, its range is between 0 and 1. However, the valid range in hypothesis testing is between 0 and 0.5, because $1 - p_{i,j}$ is the same as $p_{j,i}$,

the *p*-value of the opposite testing in Eq. (7) to verify $\mu_i > \mu_j$ (i.e., $H_0: \mu_i < \mu_j, \underline{H_A: \mu_i > \mu_j}$). Thus, $p_{i,j} = p_{j,i} = 0.5$ indicates the most uncertain situation where $\mu_i < \mu_j$ or $\mu_i > \mu_j$ cannot be determined based on the observed results [16].

Using the meaning of *p*-value in statistical hypothesis testing, the uncertainty is defined as a combination of *p*-values to evaluate the degree to which the current simulation results of each candidate can be significant evidence to verify Eq. (6). Denote that x_b is a candidate belonging to $\bar{S}_B$. All of the relationships in which x_b is involved among the relationships in Eq. (6) are

$$\left(\underline{\mu_b} < \mu_{j^{m+1}}\right) \wedge \left(\underline{\mu_b} < \mu_{j^{m+2}}\right) \wedge \cdots \wedge \left(\underline{\mu_b} < \mu_{j^k}\right) \tag{10}$$

To maximize $P\{\text{CS}\}$, the simulation results of x_b should be significant evidence to verify every relationship in Eq. (10). That is, every *p*-value of the test for each relationship (i.e., $p_{b,j^{m+1}}, \ldots, p_{b,j^k}$) has to converge on zero. However, if only a one of them has a high value even though all the other *p*-values are zero, the results of x_b cannot be evidence, and it is unclear whether x_b is an element of S_B. Accordingly, the uncertainty of x_b is defined as the maximum value among the *p*-values for verifying the relationships in Eq. (10):

$$\delta_b = \max\left(p_{b,j^{m+1}}, p_{b,j^{m+2}}, \ldots, p_{b,j^k}\right) \tag{11}$$

Because $\bar{\mu}_b < \bar{\mu}_{j^{m+1}}, \ldots, \bar{\mu}_{j^k}$, the range of δ_b is between 0 and 0.5. In common with the *p*-value, δ_b indicates the statistical significance of the observed results of x_b to verify Eq. (10). As δ_b approaches zero, the results of x_b become significant evidence that provides statistical support for Eq. (10) being true. Conversely, as δ_b approaches 0.5, it indicates an uncertain situation in which a decision cannot be made based on the current results. Meanwhile, a value of δ_b close to 1 indicates the results offer significant evidence that verifies at least one or more relationships in Eq. (10) are false; thus, it is statistically confirmed that Eq. (10) is false. This represents the proper definition of the uncertainty based on the *p*-value.

Strictly speaking, δ_b indicates the statistical significance of the results not only for x_b, but also for another candidate among $x_{j^{m+1}}, \ldots, x_{j^k}$, depending on the nature of *p*-value. However, except in some extreme cases, this does not significantly affect the allocation of Δ based on the uncertainty of each candidate. This is because each relationship in Eq. (6) is considered simultaneously in the two candidates involved in this relationship when calculating their uncertainty. In other words, if the *p*-value of any relationship in Eq. (6) is high, then such a high value is reflected in the uncertainty of both candidates included in this relationship; thus, these two candidates can be assigned at least the same number of further replications. If we can allocate relatively more replications to the candidate that has a greater influence on this high *p*-value between both candidates, then this may be a better way. However, this approach complicates the allocation method and demands additional calculations. In addition, if Δ is small, then the allocation differences will become insignificant. In this paper, we aim to develop a simple and practical allocation means to improve the simulation efficiency significantly in the identification of the best and worst subsets; thus, we neglect this issue in the proposed method.

Like δ_b, the uncertainty of the other candidates can be defined as follows. Denote that x_w is a candidate belonging to $\bar{S}_W$. Then, all of the relationships in which x_w is involved among the relationships in Eq. (6) are as below:

$$\left(\mu_{j^1} < \underline{\mu_w}\right) \wedge \left(\mu_{j^2} < \underline{\mu_w}\right) \wedge \cdots \wedge \left(\mu_{j^{k-n}} < \underline{\mu_w}\right) \tag{12}$$

To maximize $P\{\text{CS}\}$, the simulation results of x_w should be significant evidence to verify every relationship in Eq. (12); thus, the uncertainty of x_w is defined as the maximum value among the *p*-values for verifying these relationships:

$$\delta_w = \max\left(p_{j^1,w}, p_{j^2,w}, \ldots, p_{j^{k-n},w}\right) \tag{13}$$

Denote that x_l is the remaining candidate that does not belong to both $\bar{S}_B$ and $\bar{S}_W$. All of the relationships in which x_l is involved among the relationships in Eq. (6) are as follows:

$$\left\{\left(\mu_{b^1} < \underline{\mu_l}\right) \wedge \cdots \wedge \left(\mu_{b^m} < \underline{\mu_l}\right)\right\} \wedge \left\{\left(\underline{\mu_l} < \mu_{w^{k-n+1}}\right) \wedge \cdots \wedge \left(\underline{\mu_l} < \mu_{w^k}\right)\right\} \tag{14}$$

The left brace in Eq. (14) represents the relationships with candidates belonging to $\bar{S}_B$, whereas the right brace represents the relationships with candidates belonging to $\bar{S}_W$. To maximize $P\{\mathrm{CS}\}$, the simulation results of x_l should be significant evidence to verify every relationship in Eq. (14); thus, the uncertainty of x_l is defined as the maximum value among the p-values for verifying these relationships:

$$\delta_l = \max\left(p_{b^1,l}, \ldots, p_{b^m,l}, p_{l,w^{k-n+1}}, \ldots, p_{l,w^k}\right) \tag{15}$$

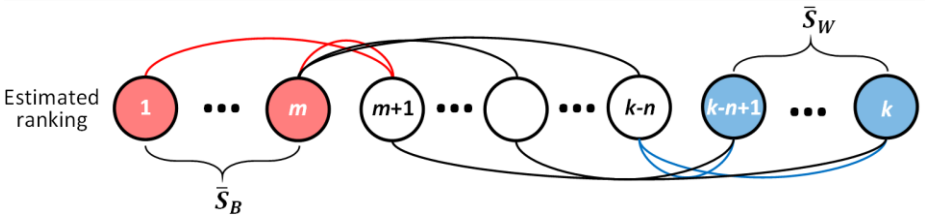

Figure 1: The figure indicates which candidates are compared to evaluate the uncertainty of k candidates that result from the approximation in Eqs. (16) and (17).

Evaluating the uncertainty of k candidates using Eqs. (11), (13), and (15) requires calculating the p-values for $(k-m)m + (k-m-n)n$ relationships in Eq. (6), which has $O(k^2)$ computational complexity in the worst case (e.g., $m = n = k/2$). As shown in Eq. (8), the p-value increases to 0.5 when the difference between $\bar{\mu}_i$ and $\bar{\mu}_j$ decreases. In many cases (e.g., s_i^2 is similar among candidates, or N_i is large enough), δ_b is the same as $p_{b,j^{m+1}}$ in Eq. (11) because $\bar{\mu}_{j^{m+1}}$ is closest to $\bar{\mu}_b$. Similarly, δ_w is the same as $p_{j^{k-n},w}$ in Eq. (13) because $\bar{\mu}_{j^{k-n}}$ is closest to $\bar{\mu}_w$. Accordingly, Eqs. (11) and (13) can be approximated as follows:

$$\delta_b \approx p_{b,j^{m+1}} \text{ and } \delta_w \approx p_{j^{k-n},w} \tag{16}$$

In the case of x_l, Eq. (15) can be divided as below according to Eq. (14):

$$\delta_l = \max\left(\max\left(p_{b^1,l}, \ldots, p_{b^m,l}\right), \max\left(p_{l,w^{k-n+1}}, \ldots, p_{l,w^k}\right)\right) \approx \max\left(p_{b^m,l}, p_{l,w^{k-n+1}}\right) \tag{17}$$

Because $\bar{\mu}_{b^m}$ is closest to $\bar{\mu}_l$ among the candidates in $\bar{S}_B$, the left maximum function in Eq. (17) can be approximated as $p_{b^m,l}$. The right maximum function also can be approximated as $p_{l,w^{k-n+1}}$ because $\bar{\mu}_{w^{k-n+1}}$ is closest to $\bar{\mu}_l$ among the candidates in $\bar{S}_W$. As a result, δ_i is defined as the maximum value among $p_{b^m,l}$ and $p_{l,w^{k-n+1}}$. This approximation significantly reduces the complexity for evaluating uncertainty from $O(k^2)$ to $O(k)$, while keeping the feature of the uncertainty. Here, the complexity of $O(k)$ is the same level as the existing methods such as OCBAmn. In addition, because of the approximation, the uncertainty of each candidate is evaluated through comparison with the marginal candidates of each group (e.g., x_m, x_{m+1}, x_{k-n}, and x_{k-n+1}), which is described in Fig. 1. This can imply the approximation is reasonable because these marginal candidates are criteria for distinguishing S_B and S_W.

To maximize $P\{\mathrm{CS}\}$ within T, Δ replications have to be allocated so that simulation results of all candidates are significant evidence to verify Eq. (6). That is, depending on the uncertainty of each candidate evaluated with Eqs. (16) and (17), candidates that have relatively high uncertainty should be allocated additional replications to update their deficient simulation results. However, candidates that have relatively low uncertainty also need to be considered in the allocation of Δ replications due to the type I error involved in the uncertainty. This error often occurs in a high-noise situation, specifically, when the sample mean of simulation outputs is very biased due to a small number of samples and a large variance of the output. For example,

after conducting initial n_0 replications for each candidate, the sample mean of a candidate that does not belong to S_B and S_W is unfortunately very biased (i.e., $\bar{\mu}_i \ll \mu_i$), so this candidate is included in $\bar{S}_B$ and its sample mean is even smaller than that of the marginal candidate in $\bar{S}_B$ (i.e., $\bar{\mu}_i \ll \bar{\mu}_{b^m}$). In this case, the uncertainty of this candidate is small by Eq. (16), but this is an error and it causes incorrect allocation.

To correct the error, additional replications should be allocated to this candidate to improve its biased results. However, although the allocation of Δ replications iterates continuously, it cannot be given any replications for a while due to its lower uncertainty. As shown in Eq. (8), the uncertainty becomes lower as the number of samples N_i increases; thus, as the allocation of Δ replications is repeated, the uncertainty of the candidate that has not been allocated further replications becomes relatively higher than that of the other candidates that have been allocated replications, and eventually, it can receive additional replications. However, in the meantime, a large amount of simulation budget is wasted by incorrect allocation, and $\bar{S}_B$ and $\bar{S}_W$ have not been correctly selected. The higher the noise, the more often such type I error occurs; thus, quickly allocating further replications to these candidates with the error to enhance the robustness to noise is necessary. Fortunately, the simulation results of these candidates are very biased based on a small number of samples, so even with a few normal samples, the results are greatly improved and the error can be corrected. That is, a few further replications should be assigned to the candidates that have relatively low uncertainty to increase the robustness.

In summary, to increase $P\{\mathrm{CS}\}$, many further replications should be allocated to candidates that have relatively high uncertainty depending on the meaning of the uncertainty; however, candidates of which the uncertainty is relatively low also simultaneously need a few additional replications to enhance the robustness in a high-noise situation. To achieve this, the proposed method applies the following heuristic policy [8] for allocating Δ replications to k candidates based on their uncertainty:

$$a_i / a_j = \left(\delta_i / \delta_j\right)^C \text{ where } i, j \in \{1,\ldots,k\} \text{ and } \Delta = \sum_{i=1}^{k} a_i \tag{18}$$

a_i (or a_j) is the number of allocated further replications. Although a_i should be an integer, it is likely will not be, according to Eq. (18). If a reasonable conversion method is used, then neither method has a significant effect on $P\{\mathrm{CS}\}$ [7]. In the proposed method, we used the round function to convert a_i to an integer (see Line 8 in Algorithm 1).

In the policy of Eq. (18), C is the parameter to adjust the allocation based on the uncertainty. Its range is between zero and infinity. As C decreases, candidates that have relatively low uncertainty can be allocated more replications than before, and if C is zero, then all candidates receive the same replications regardless of uncertainty. On the other hand, as C increases, Δ replications are allocated to only a few candidates of which the uncertainty is relatively high, and if C goes to infinity, then the candidate with the maximum uncertainty takes it all. The optimal setting of C to maximize $P\{\mathrm{CS}\}$ within T is problem-specific and related to the noise level of the problem. Here, the noise level does not simply mean the absolute value of variance in simulation output, but the relative difference with the mean performance of candidates. In a high-noise environment, a small value such as 0.5 can be a good choice to deal with frequently occurring type I errors effectively. However, these errors do not occur well when the noise level is low; thus, it can be more effective to allocate more replications to candidates with relatively high uncertainty via a high value of C, such as 2.

However, because information on the noise level of the problem is not typically given, we recommend 1 as the default value of C in the proposed method. This is not always optimal but exhibits sufficiently superior performance compared to the existing methods (see Table II, Table III, and Fig. 2). In addition, the consistency in the uncertainty evaluation of the proposed method can be a basis for this recommendation. When n is zero, the problem in Eq. (5) becomes

the same as the problem of identifying only the best subset; in this case, Eqs. (16) and (17) are also identical to the uncertainty evaluation of the existing method for selecting the best subset based on the UE method [17]. In addition, if both m and n are one, then the problem is equal to the problem of selecting the best and worst candidates; in this case, except for the approximation, the uncertainty evaluation of the proposed method is the same as that of the existing method for solving this problem based on the UE method [10]. When m is 1 and n is zero, the problem becomes identical to the problem of finding only the best candidate; in this case, except for the approximation, the evaluation of the proposed method is equal to that of the UE method [8]. Because these existing methods utilize the allocation policy of Eq. (18) with $C = 1$ depending on the empirical results (e.g., simulation-based optimization), this recommended default value can be a good choice in terms of consistency.

Algorithm 1 presents the proposed method for identifying the best and worst subsets, which includes the initial allocation of n_0 replications for every candidate to obtain the minimum simulation results, and the iterative allocation of Δ replications based on the evaluated uncertainty using Eqs. (16) to (18). Similar to the parameter C, the optimal setting of n_0 and Δ is also problem-specific. In general, the recommended value in the literature is between 5 and 20 for n_0 [18], and between $0.5k$ and $0.05T$ for Δ [8]. However, if T is large enough in a high-noise situation, increasing n_0 and Δ higher than the recommended value may be effective. A large value of n_0 reduces the biased sample mean caused by a small number of simulation output samples and improves the quality of initial simulation results of each candidate. In addition, a large Δ value increases the likelihood that candidates with relatively low uncertainty will be allocated a few further replications to correct the type I error quicker.

Algorithm 1 Identify the best and worst subsets (m, n) within a given simulation budget T

Control parameters: n_0, Δ, and C (default value is 1)

Procedure:

simulate n_0 times for each x_i, $i \in \{1, \dots, k\}$
update $\bar{\mu}_i$, s_i, and N_i for $\forall i \in \{1, \dots, k\}$
select $\bar{S}_B$ and $\bar{S}_W$ with Eqs. (3) and (4)
while $\sum_{i=1}^{k} N_i < T$ **do**
  set $\Delta \leftarrow \min\left(T - \sum_{i=1}^{k} N_i, \Delta\right)$
  evaluate δ_b, δ_w, and δ_l with Eqs. (16) and (17)
  calculate a_i for $\forall i \in \{1, \dots, k\}$ with Eq. (18)
  simulate round$(a_i)^*$ times for each x_i, $i \in \{1, \dots, k\}$
  update $\bar{\mu}_i$, s_i, and N_i for $\forall i \in \{1, \dots, k\}$
  select $\bar{S}_B$ and $\bar{S}_W$ with Eqs. (3) and (4)
end while
return $\bar{S}_B$ and $\bar{S}_W$

* The round function is used to convert a_i to an integer. If the rounding errors are positive, they are scattered and added to several candidates that have not been assigned further replications; otherwise, they are subtracted from some candidates that have been assigned a lot.

4. EXPERIMENTS

In this section, we exhibit experimental results to demonstrate the improved efficiency and robustness of the proposed method compared to the existing methods, such as the OCBAmn [13], hUEm, hOCBAm+, and Equal methods. Here, hUEm and hOCBAm+ methods are variations of UEm [17] and OCBAm+ [12] methods, which are methods for identifying the best subset, respectively. They use their original method to find the best subset with half of the given T and find the worst subset with the other half. The comparison with these two methods will indicate the necessity of the proposed method. The equal method is the most basic method of allocating T/k replications equally to all candidates.

Table I: Numerical benchmark problems.

Problem	k	Output distribution	Output distribution (L)
Equal variance (EV)	50	$N(i, 10^2)$	$N(\mathbf{0.5} \times i, 10^2)$
Decreasing variances (DV)	50	$N(i, ((k+1-i)/3)^2)$	$N(\mathbf{0.5} \times i, ((k+1-i)/3)^2)$
Increasing variances (IV)	50	$N(i, (i/3)^2)$	$N(\mathbf{0.5} \times i, (i/3)^2)$
Convex increasing mean (CVM)	50	$N\left(k-\sqrt{k}\sqrt{k-i}, 5^2\right)$	$N\left(\mathbf{0.5} \times \left(k-\sqrt{k}\sqrt{k-i}\right), 5^2\right)$
Concave increasing mean (CCM)	50	$N\left(k-\left((k-i)/\sqrt{k}\right)^2, 3^2\right)$	$N\left(\mathbf{0.5} \times \left(k-\left((k-i)/\sqrt{k}\right)^2\right), 3^2\right)$

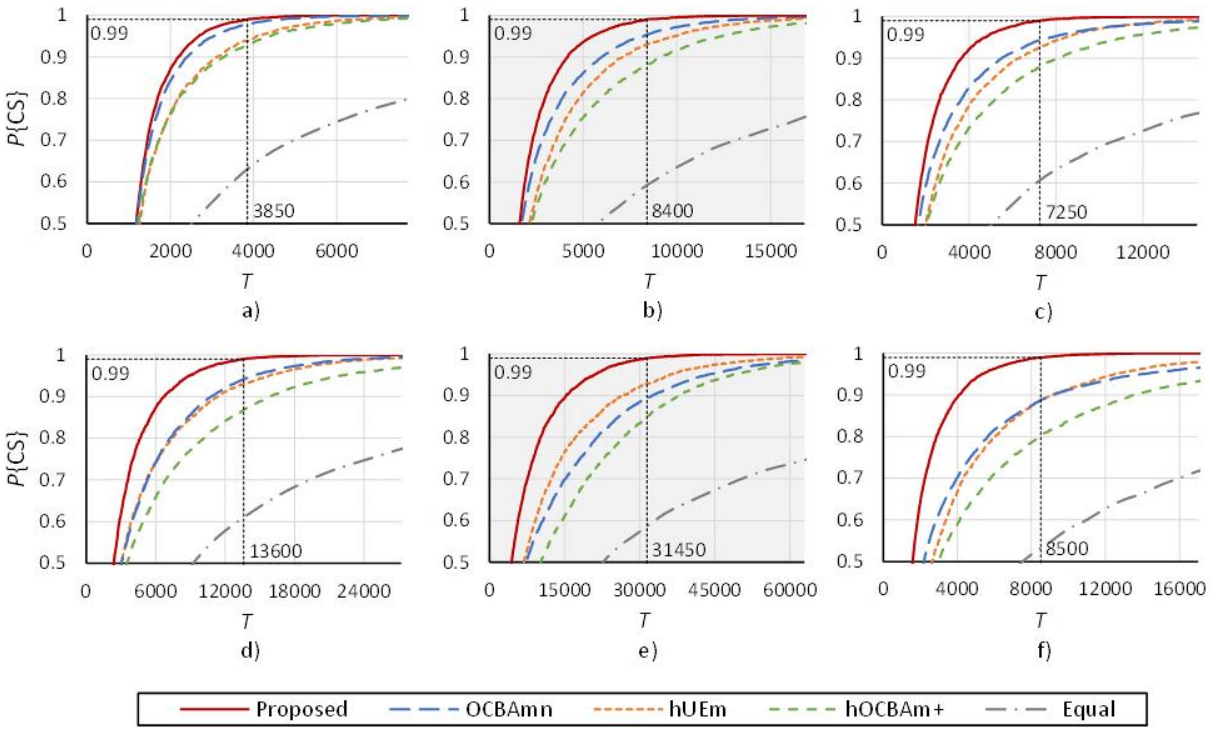

Figure 2: The estimated value of $P\{\mathrm{CS}\}$ versus T for the selected benchmark problems: a) CVM[5, 5], b) DV[4, 6], c) CCM[4, 6], d) CVM(L)[5, 5], e) DV(L)[4, 6], f) CCM[5, 5], where the numbers in square brackets indicate m and n.

We applied the five benchmark problems that were used to evaluate the OCBAmn method, and Table I summaries them. Each problem has 50 candidates, and their mean performance and variance are defined in Table I. The five problems have two versions. The (L) version has a relatively large noise level compared to its original version by reducing the mean performance of candidates while maintaining the variance. For a total of 10 problems, we considered two cases: (1) $m = n = 5$ and (2) $m = 4$, $n = 6$. Except for the equal method, all methods have the same setting (i.e., $n_0 = 20$ and $\Delta = 50$) for a fair comparison, where this setting was used to evaluate the OCBAmn method. In addition, C in the proposed method was set to the default value of 1. For each problem and case, $P\{\mathrm{CS}\}$ according to T was estimated through 10,000 independently repeated experiments for each method (Fig. 2 presents some results). Table II and Table III indicates simulation budget required to identify the best and worst subsets correctly (i.e., to achieve $P\{\mathrm{CS}\}$ of 0.99). The experimental results clearly demonstrate the improved efficiency and high robustness of the proposed method over the existing methods.

The comparison results with hUEm and hOCBAm+ show the necessity of the proposed method for identifying the best and worst subsets simultaneously. In EV and EV(L) problems, all candidates have the same variance, and the difference between their mean performance is constant; thus, the cost to identify the best and worst subsets is almost the same. In this case, the efficiency gap between the proposed method and these two methods is not (relatively) large.

However, in the other problems, the cost is not the same due to the non-uniform difference of the mean performance or unequal variances. As shown in Table II and Table III, the efficiency gap is significantly large in these problems. In addition, as shown in Table I, the CCM problem requires a higher simulation budget to find the worst subset than the it does for best subset. The efficiency gap in this problem is much larger when $m = 4$ and $n = 6$ than it is when $m = n = 5$, as described in Figs. 2 c and 2 f. Because such a situation in which the cost for selecting the best and worst subsets is different is common in real problems, practitioners need the proposed method to identify both subsets efficiently.

Table II: Simulation budget T required to achieve $P\{\text{CS}\}$ of 0.99 for benchmarks ($m = 5$, $n = 5$).

Problem	Proposed	OCBAmn	hUEm	hOCBAm+	Equal
EV	7,000	10,150 **(1.45)***	8,500 **(1.21)**	10,000 **(1.43)**	66,500 **(9.50)**
EV(L)	30,850	54,050 **(1.75)**	37,600 **(1.22)**	51,300 **(1.66)**	272,000 **(8.82)**
DV	8,350	12,600 **(1.51)**	15,200 **(1.82)**	18,850 **(2.26)**	123,000 **(14.73)**
DV(L)	32,600	66,500 **(2.04)**	59,400 **(1.82)**	69,350 **(2.13)**	485,000 **(14.88)**
IV	8,500	12,700 **(1.49)**	15,550 **(1.83)**	19,500 **(2.29)**	117,000 **(13.76)**
IV(L)	36,000	68,450 **(1.90)**	60,700 **(1.69)**	70,150 **(1.95)**	500,000 **(13.89)**
CVM	3,850	4,700 **(1.22)**	6,600 **(1.71)**	7,200 **(1.87)**	49,000 **(12.73)**
CVM(L)	13,600	23,350 **(1.72)**	25,300 **(1.86)**	37,700 **(2.77)**	190,000 **(13.97)**
CCM	8,500	33,850 **(3.98)**	20,700 **(2.44)**	39,400 **(4.64)**	145,000 **(17.06)**
CCM(L)	49,450	227,400 **(4.60)**	87,150 **(1.76)**	≫300,000 [0.9266]**	620,000 **(12.54)**

* The number in brackets denotes the ratio of T required in the proposed method and that required in the other method (i.e., $T_{\text{other}}/T_{\text{proposed}}$).

** Even if the simulation budget is sufficient (i.e., $T = 300{,}000$), $P\{\text{CS}\}$ only reaches the value in the square bracket (i.e., more budget is required to achieve $P\{\text{CS}\}$ of 0.99).

Table III: Simulation budget T required to achieve $P\{\text{CS}\}$ of 0.99 for benchmarks ($m = 4$, $n = 6$).

Problem	Proposed	OCBAmn	hUEm	hOCBAm+	Equal
EV	6,900	10,150 **(1.47)***	8,400 **(1.22)**	9,850 **(1.43)**	67,500 **(9.78)**
EV(L)	29,900	52,200 **(1.75)**	36,550 **(1.22)**	51,850 **(1.73)**	268,000 **(8.96)**
DV	8,400	13,400 **(1.60)**	16,000 **(1.90)**	20,000 **(2.38)**	129,000 **(15.36)**
DV(L)	31,450	69,600 **(2.21)**	58,350 **(1.86)**	74,400 **(2.37)**	485,000 **(15.42)**
IV	8,100	12,450 **(1.54)**	15,000 **(1.85)**	16,850 **(2.08)**	121,000 **(14.94)**
IV(L)	34,950	66,850 **(1.91)**	58,700 **(1.68)**	62,950 **(1.80)**	483,000 **(13.82)**
CVM	3,950	4,700 **(1.19)**	6,550 **(1.66)**	7,250 **(1.84)**	47,000 **(11.90)**
CVM(L)	13,900	24,350 **(1.75)**	23,850 **(1.72)**	38,250 **(2.75)**	195,000 **(14.03)**
CCM	7,250	15,600 **(2.15)**	13,350 **(1.84)**	20,250 **(2.79)**	100,000 **(13.79)**
CCM(L)	31,900	100,050 **(3.14)**	57,350 **(1.80)**	>300,000 [0.9837]**	388,000 **(12.16)**

* The number in brackets denotes the ratio of T required in the proposed method and that required in the other method (i.e., $T_{\text{other}}/T_{\text{proposed}}$).

** Even if the simulation budget is sufficient (i.e., $T = 300{,}000$), $P\{\text{CS}\}$ only reaches the value in the square bracket (i.e., more budget is required to achieve $P\{\text{CS}\}$ of 0.99).

The experimental results for the (L) version in Table II and Table III demonstrate the high robustness of the proposed method compared to the other methods, in particular OCBAmn method. When comparing Figs. 2 a and 2 b with Figs. 2 d and 2 e, respectively, the efficiency gap between the proposed and OCBAmn methods is much larger in the (L) version than the original version of the problems. This is because the OCBAmn method allocates Δ replications with only the values of the sample mean and sample variance of each candidate, whereas the proposed method additionally utilizes the number of samples (i.e., N_i). Namely, the proposed method considers the precision of the sample mean defined by $s_i/\sqrt{N_i}$ as well as the value of the sample mean. Suppose a candidate in S_B has a biased value of the sample mean that reduces $P\{\text{CS}\}$ (i.e., it does not belong to $\bar{S}_B$) in a high-noise situation. Although this candidate needs further replications to fix its distorted sample mean, the OCBAmn method does not allocate the

replications based on the biased value (i.e., OCBAmn does not consider this candidate as an element of S_B depending on the distorted value). The proposed method also does not allocate further replications to this candidate due to the type I error. Here, the difference in robustness between both methods is attributed to how quickly and surely further replications are allocated to this candidate. In the proposed method, as the allocation of Δ replications iterates, the uncertainty of this candidate that has not been allocated further replications becomes relatively higher compared to the others (i.e., the precision of the sample mean of this candidate is relatively lower compared to the others), and a few replications may soon be allocated. On the other hand, because the OCBAmn method considers only the value of the sample mean in the allocation, whether the OCBAmn method can allocate further replications to a candidate whose sample mean is no longer being updated is highly uncertain. That is, the possibility of the further allocation depends on changes in the sample mean and variance of the other candidates. Thus, sometimes even if T is large, this candidate will eventually not receive any further replications, which in turn lowers the efficiency of OCBAmn in high-noise situations.

5. CONCLUSION

In this paper, we proposed a robust method for identifying the best and worst subsets from k candidates in stochastic simulation. To maximize $P\{\mathrm{CS}\}$ within a limited simulation budget, the proposed method efficiently allocates the budget to candidates based on the UE method. That is, it defined the uncertainty to evaluate the statistical significance of each candidate's simulation results using the statistical hypothesis test and p-value. Then, in the iterative allocation procedure, small further simulation replications are allocated in proportion to the calculated uncertainty of each candidate and the simulation results are updated. By doing so, each candidate's results become significant evidence that verifies the correct selection of the best and worst subsets. Experimental results on benchmark problems demonstrated that the proposed method could achieve significantly higher $P\{\mathrm{CS}\}$ than the existing methods in a high-noise situation. Such high robustness is because the proposed method can fix the distorted sample means quickly by considering the precision of the sample mean when allocating further replications based on the uncertainty.

As mentioned earlier, the proposed method can be applied to best-worst scaling problems, multiple-criteria decision-making, population-based search algorithms, etc., where the best and worst subsets should be identified in stochastic environments. In addition, in simulation experiments, it allows the practitioners to make final decisions by considering qualitative criteria neglected by simulation. In particular, its excellent efficiency in high-noise situations can satisfy the time constraints for simulation experiments on complex and costly modern simulation models such as digital twins. Meanwhile, the high robustness to noise acts as a disadvantage in low-noise situations where the distorted sample means do not often occur. However, this may not be an important issue because the number of simulation replications itself required to identify the best and worst subsets is small when the noise level is low. In addition, because most practical simulation models have large noise to capture the high levels of uncertainty in real systems, the proposed method can be still effective.

The proposed method assumes the simulation output sample in a replication follows an independent and identically distributed normal distribution, which is a rather strong assumption in some practical situations. To enhance practicality, future research can be devoted to considering various distributions in simulation output and the correlation among outputs.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea Government (Ministry of Science and ICT) under Grant 2019R1G1A1098951.

REFERENCES

[1] Amorim, G. A.; Lopes, L. A. S.; Silva Jr., O. S. (2020). Discrete event-based railway simulation model for eco-efficiency evaluation, *International Journal of Simulation Modelling*, Vol. 19, No. 3, 375-386, doi:10.2507/IJSIMM19-3-517

[2] Ojstersek, R.; Acko, B.; Buchmeister, B. (2020). Simulation study of a flexible manufacturing system regarding sustainability, *International Journal of Simulation Modelling*, Vol. 19, No. 1, 65-76, doi:10.2507/IJSIMM19-1-502

[3] Freile, A. J.; Mula, J.; Campuzano-Bolarin, F. (2020). Integrating inventory and transport capacity planning in a food supply chain, *International Journal of Simulation Modelling*, Vol. 19, No. 3, 434-445, doi:10.2507/IJSIMM19-3-523

[4] Xu, J.; Huang, E.; Chen, C.-H.; Lee, L. H. (2015). Simulation optimization: a review and exploration in the new era of cloud computing and big data, *Asia-Pacific Journal of Operational Research*, Vol. 32, No. 3, Paper 1550019, 34 pages, doi:10.1142/S0217595915500190

[5] Kim, S.-H.; Nelson, B. L. (2006). Selecting the best system, Henderson, S. G.; Nelson, B. L. (Eds.), *Handbooks in Operations Research and Management Science: Simulation*, Vol. 13, Elsevier, Amsterdam, 501-534

[6] Chick, S. E.; Branke, J.; Schmidt, C. (2010). Sequential sampling to myopically maximize the expected value of information, *INFORMS Journal on Computing*, Vol. 22, No. 1, 71-80, doi:10.1287/ijoc.1090.0327

[7] Chen, C.-H.; Lee, L. H. (2011). *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*, World Scientific Publishing, Singapore

[8] Choi, S. H.; Kim, T. G. (2018). Efficient ranking and selection for stochastic simulation model based on hypothesis test, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 48, No. 9, 1555-1565, doi:10.1109/TSMC.2017.2679192

[9] Flynn, T. N.; Louviere, J. J.; Peters, T. J.; Coast, J. (2007). Best–worst scaling: what it can do for health care research and how to do it, *Journal of Health Economics*, Vol. 26, No. 1, 171-189, doi:10.1016/j.jhealeco.2006.04.002

[10] Choi, S. H.; Choi, C. (2020). A simulation budget allocation procedure for finding both extreme designs simultaneously in discrete-event simulation, *IEEE Access*, Vol. 8, 93978-93986, doi:10.1109/ACCESS.2020.2995196

[11] Zhang, J.; Zhang, L.; Wang, C.; Zhou, M. (2017). Approximately optimal computing budget allocation for selection of the best and worst designs, *IEEE Transactions on Automatic Control*, Vol. 62, No. 7, 3249-3261, doi:10.1109/TAC.2016.2628158

[12] Zhang, S.; Lee, L. H.; Chew, E. P.; Xu, J.; Chen, C.-H. (2016). A simulation budget allocation procedure for enhancing the efficiency of optimal subset selection, *IEEE Transactions on Automatic Control*, Vol. 61, No. 1, 62-75, doi:10.1109/TAC.2015.2423832

[13] Xiao, H.; Gao, S.; Lee, L. H. (2017). Simulation budget allocation for simultaneously selecting the best and worst subsets, *Automatica*, Vol. 84, 117-127, doi:10.1016/j.automatica.2017.07.006

[14] Goldberg, D. E.; Deb, K. (1991). A comparative analysis of selection schemes used in genetic algorithms, Rawlins, G. J. E. (Ed.), *Foundations of Genetic Algorithms*, Vol. 1, Elsevier, Amsterdam, 69-93

[15] Zhang, J.; Li, Z.; Wang, C.; Zang, D.; Zhou, M. (2017). Approximate simulation budget allocation for subset ranking, *IEEE Transactions on Control Systems Technology*, Vol. 25, No. 1, 358-365, doi:10.1109/TCST.2016.2539329

[16] Hayter, A. (2012). *Probability and Statistics for Engineers and Scientists*, 4th edition, Cengage Learning, Boston

[17] Choi, S. H.; Kim, T. G. (2018). Optimal subset selection of stochastic model using statistical hypothesis test, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 48, No. 4, 557-564, doi:10.1109/TSMC.2016.2608982

[18] Bechhofer, R. E.; Santner, T. J.; Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*, John Wiley & Sons, New York