Information Technology and Management Science

ISSN 2255-9094 (online)
ISSN 2255-9086 (print)
December 2018, vol. 21, pp. 86–92
doi: 10.7250/itms-2018-0014
https://itms-journals.rtu.lv

Approach to Integration of Data Mining Techniques in Simulation Results Analysis

Irīna Šitova[1], Jeļena Pečerska[2]
[1,2] Riga Technical University, Riga, Latvia

***Abstract* – The research is carried out in the area of analysis of simulation results by using data mining techniques. The goal of the research is to explore the applicability of data mining techniques in the area of simulation results analysis, to offer an application scheme of data mining techniques in the analysis of simulation results, as well as to demonstrate the usage of these techniques in the analysis of experimental data. As a result of the theoretical study, an approach is proposed, consisting of two stages and combining the fundamental techniques of data farming and knowledge discovery. A variety of data mining techniques, such as correlation analysis, clustering and several visualization mechanisms of results, are used for knowledge discovery. The proposed approach is applied to the analysis of experimental data. The performance of a queueing system is analysed, and knowledge and decision rules are obtained from simulation results.**

***Keywords* – Data mining, discrete-event system simulation, queueing system, simulation results analysis.**

I. Introduction

Simulation and data mining are two technologies; each helps solve different problems in various areas of life. A simulation is the imitation of a complex system that gives a possibility to display and describe the behaviour of a system in a detailed way [1]. In the process of simulation, random factors influencing the system are taken into account, as well as their changes over time. Researchers obtain information about the system under consideration after experiments with a verified, calibrated and validated model [2]. In the present paper, the discussion is limited to the results obtained from a discrete-event system (DES) simulation models, which are most effectively used for the analysis of dynamics of complex artificial material systems.

The traditional simulation report is generalization of the output statistics after running the simulation model. The output statistics is based on the results of a planned experiment or series of experiments. “Output analysis is the examination of data generated by simulation” [3]. The output statistics is the main “trophy” of the researcher simulationist and allows understanding the behaviour of the system, to formulate forecasts, to compare alternatives or to solve the optimization tasks of system parameters. The question under consideration is: How should one perform data analysis and interpret the results? The traditional approach to simulation output analysis implies replication design, estimation of performance indicators and analysis of system behaviour, based on these estimations [4]. However, this final analysis cannot be reduced to statistical methods only.

Various researchers [5]–[10] are investigating simulation results combining the statistical analysis and data mining. Data mining is described as the process, during which previously unknown, non-trivial, practically useful and accessible knowledge is obtained from the raw data [11]. This technology is capable of finding significant linkages in big data arrays and identify behavioural patterns in order to help people make reasonable decisions.

As a result of data mining techniques application in simulation results analysis, the knowledge is obtained, which increases the simulation efficiency. It provides a more versatile output analysis and deals with a potentially huge amount of simulation output data. The general scheme of integration of two powerful technologies is shown in Fig. 1.

The authors provide a review of approaches in the area of analysis of simulation results by using data mining techniques. The goal of the present research is to explore the applicability of data mining techniques in the area of simulation result analysis and to introduce an application scheme of data mining techniques for the analysis of simulation results. As a result of the theoretical study, a two-stage approach is formulated, combining the fundamental principles of data farming and knowledge discovery. A variety of data mining techniques, including correlation analysis, clustering and several visualization mechanisms of results, are used for the knowledge discovery. The developed approach is applied to the analysis of experimental data of a simple DES simulation model.

The authors hypothesise that data mining techniques may provide a better interpretation of simulation output as well as visualization of outputs. Another issue is to make sure that data mining reveals not only trivial knowledge from simulation output. Finally, knowledge and decision rules are obtained from simulation results coupled with the relevant visualization.

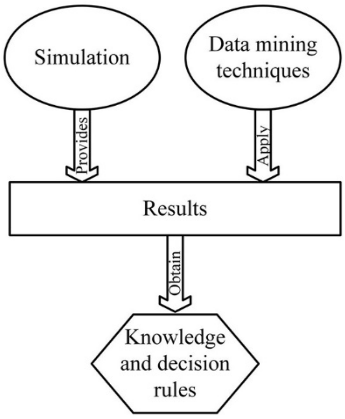


Fig. 1. The scheme of integration of data mining techniques in the simulation results analysis.

II. Related Work

A. Heuristic-Based Searches with Simulation Models, 2001

One of the pioneers in the field of integrating data mining techniques in simulation results analysis were Brady and Bowden [5]. They used heuristic-based searches with simulation models for solving one of the most important simulation challenges – selecting the optimal set of input variables and their values. There is a rule – a compromise between adding all important input variables and not overloading the model must be found. The provided solution is named an 'external' optimization, as soon as an input variable determining process is made outside of the simulation model. The method can provide 'better' answers than trial-and-error methods, when the selection of the optimal set of input variables is made based on a large number of experiments. However, Brady and Bowden mention the disadvantage of the proposed method, which is associated with the fact that people, who may not be accurate enough or even wrong, implement the most important phase – the selection of decision variables for optimization. As a result, the method is excessively dependent on decision makers that define input variables.

B. A New Form of Computer Simulation Output, 2005

In contrast to [5], Brady and Yelling in [6] developed an internal approach for selecting optimization variables that analysed the dynamics of input variable interaction within the simulation model. The aim of the research was to create a method that used only one replication results of the simulation model to order the elements according to their importance. Simulation model elements, including resources, entities, statuses, and their relationship display the logic of a model. During one replication, the information that appears in the trace file is analysed. Each simulation model element has a code or a keyword for easier information interpretation from a trace file. The frequency analyser program evaluates the frequency of keyword appearance in a trace file. Then, considering the information obtained, a correlation analysis is performed using the cosine method. The method described in [6] was applied to the practical problem, where the information obtained was considered non-trivial. This information was transformed into useful knowledge that helped improve the semiconductor manufacturing process. To conclude, it is appropriate to use the proposed method when the simulation model consists of a big number of elements and it is necessary to know their interconnectedness.

C. Combined Techniques Including Simulation, Data Mining and Knowledge Discovery, 2006

The method described in [10] combines simulation, data mining and knowledge discovery techniques for optimising aircraft engine (both short-term and long-term) life cycle costs (LCC). The method is based on the following algorithm: the obtained data on operational functionality and costs from the simulation model are mined in order to discover the parametric relations that best describe the LCC value changes for each accepted strategic decision.

Painter et al. in [10] use several data mining methods, such as linear regression, clustering, and classification, to determine significant cost drivers. Two software tools are used: Arena as a simulation tool and the PDP as a data mining tool.

To test the effectiveness of the method, a case study was carried out. First, a linear regression model was constructed to find the values that affected the LCC increase.

Then for the analysis of the parameters that affected the LCC value, the methods of classification were used. The classification was performed using the decision tree-based algorithm – CART. Finally, the simulation results were analysed using a clustering algorithm – K-means algorithm. The main result of the research [10] was the proof that there was a possibility of using data mining methods to identify and describe cost drivers, obtaining them from the simulation results.

D. Integrating Data Analytics and Simulation Methods, 2015

Another research in the area [9] provides a methodology, according to which, values that most effectively increase the investigated system performance are extracted from data using data analytics methods. These values are used to prepare the appropriate simulation input data for execution of scenarios. System optimization is performed based on simulation output data.

Two features distinguish the described methodology [9] from traditional simulation and optimization approaches: a) the input data which are collected from various systems using smart devices (e.g., sensors) are large in scale, include different values and they are constantly replenished; b) usage of data mining methods, including association and classification methods, to identify more relevant variables that are related to specific performance indicators.

Kibira et al. in [9] demonstrate the application of the methodology for the metal product manufacturing system model. As a result, a specific process plan is defined, which optimises production costs and the methodology that integrates data mining, simulation and optimization to help make more constructive decisions.

E. Further Development of Data Mining Based Approaches, Actualities

Feldkamp et al. in [7] develop and test the proposed approach by formulating an applicable management strategy in [8]. The sequential application of data mining techniques – correlation analysis, clustering, and best cluster selection – resulted in a finding of prevailing input parameter values. The obtained information is used for a decision about changing system parameter values. The previously formulated approach is successfully implemented for real-world problem solving.

There is a considerable number of articles and projects concerning the merging of the concepts of data mining techniques in simulation results analysis. However, the scheme of application of data mining technology may differ from project to project and a sustainable approach in the area has not emerged yet. Taking all these factors into account, the authors of the present article propose their own approach that is described below.

III. The Developed Approach

After analysing the reviewed scientific publications [5]–[10], as well as several books [12]–[14] devoted to simulation, data mining and integration of those technologies, the authors have developed their approach. The approach includes the analysis of simulation results using several data mining techniques in the proposed order.

Publications [7] and [8] influenced the creation of the authors' approach. In the previously mentioned publications, a method is described, where data farming and knowledge discovery are combined. The authors' approach also consists of these parts, however, unlike works [7] and [8], another combination of the implemented data mining methods is applied. In the developed approach, each of the used technologies has its own purpose:

- Data farming, the purpose of which is to get results from a simulation model that corresponds to some real process.
- Knowledge discovery, the purpose of which is to find patterns in a simulation model state variable behaviour, in order to uncover knowledge which would otherwise be hidden.

The overall scheme of the developed approach, which reflects data farming and knowledge discovery phases, is shown in Fig. 2.

The following case study tends to contribute to obtaining knowledge from simulation results by applying the data mining technique.

IV. Case Study. Queueing System Simulation and Results Analysis

A. Problem Statement

An appropriate analysis of simulation model results is one of the fundamental phases of the simulation procedure. The success of the experiment outcome depends on it. Experiment outcome that is capable of improving simulated process efficiency in real life can be called a successful one. In this case, data mining techniques are coming to help.

A simple queuing system with a queue of infinite capacity, exponential distribution of interarrival times, variable average value of this exponential distribution at different daytime, random service time, queue-length based probability of lost customers is an object of the case study. Interarrival times and service times are independent. Service time is a function of demand for product. The variable demand for product is used for evaluation of the revenue, and demand distribution is described as empirical one. Probability of lost customers is described as a function of queue length. Thus, the system performance is not stable and indicators – service quality and revenue – are conflicting. As a result, it is necessary to formulate a compromise approach to achieve the best possible performance.

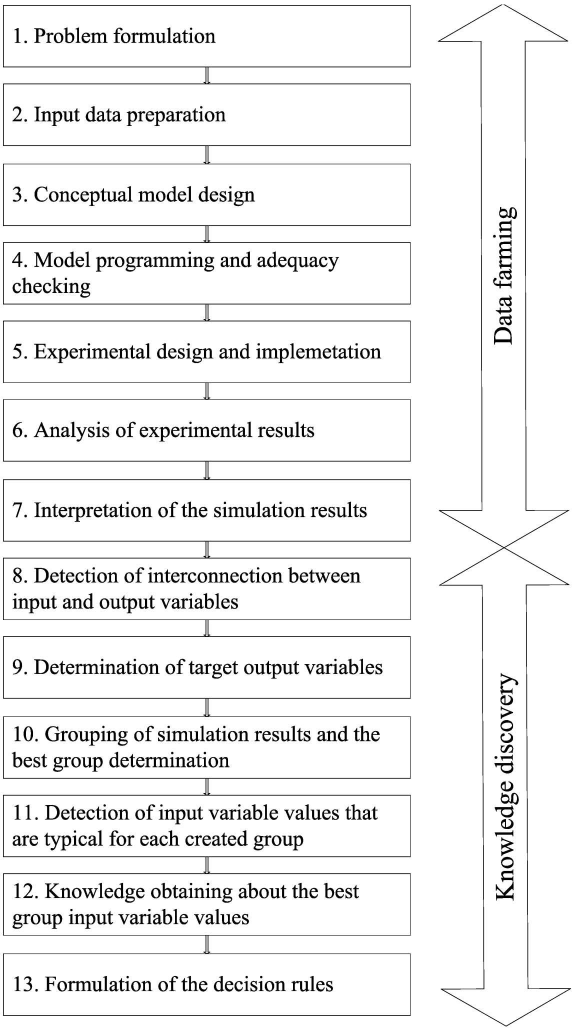

1. Problem formulation
2. Input data preparation
3. Conceptual model design
4. Model programming and adequacy checking
5. Experimental design and implemetation
6. Analysis of experimental results
7. Interpretation of the simulation results
8. Detection of interconnection between input and output variables
9. Determination of target output variables
10. Grouping of simulation results and the best group determination
11. Detection of input variable values that are typical for each created group
12. Knowledge obtaining about the best group input variable values
13. Formulation of the decision rules
Data farming
Knowledge discovery

Fig. 2. The developed approach scheme.

The detailed analytical study of system performance measures is complicated. For the purposes of performance analysis, a discrete-event simulation model of a system is created. Data mining techniques are applied to determine the parameter values for obtaining such performance.

B. The Summary of the Research Approach

To achieve the goal of the study, the process is consistently implemented as follows. Data farming realization: the simulation model of the system under consideration is constructed, simulation experiments are designed, replicated and results obtained. Knowledge discovery realization: the simulation results are analysed using data mining techniques. The domain of acceptable values of experimental factors is revealed and visualised.

C. Software Implemented for the Case Study

It can be stated that the expected performance indicators estimations are provided by any of the discrete-event simulation tools. Therefore, one of the widely used simulation packages has been applied to create a simulation model. For the analysis of simulation results, the R software environment was used as a powerful data mining tool with available visualization [15].

D. Queuing System with a Single Server, a Queue with Infinite Capacity, and Several Random Factors

The choice of the system for the case study is determined as follows: the system should be simple enough, yet the simulation results are not interpreted unambiguously –objective function alternatives of input variable sets are under consideration.

For the purposes of performance analysis, a discrete-event simulation model of a system is created, providing the estimates of relevant performance measures: statistics of customers in the system, occupation rate of the server, which part of the customer will be served, and which will not, quantity of sold goods and some other performance measures [16]. The input and output variables of the simulated model are presented in Table I. The conceptual model [17] of queueing system under consideration is shown in Fig. 3.

The goal of the case study is to find the combination of input variable values and adjustable timetables of work, providing the best values of the composite objective function.

E. Simulation Experiments and Statistical Results

To obtain data for mining, 28 model experiment scenarios have been developed with the realistic experimental factor values. Each of the scenarios has been replicated 25 times, providing 700 experiment results sets.

Fig. 4 provides an overview of the model layout, a snapshot of some simulation results and the result table used for further analysis.

F. Data Mining Technique Application in the Simulation Results Analysis

In the frame of the research, the data mining techniques have been applied to analyse corresponding data sets.

TABLE I
THE INPUT AND OUTPUT VARIABLES OF THE SIMULATION MODEL

Input variables	Output variables
Customers' arrival rate	*Number of customers entered entry point*
Probability values for decision making	*Number of unserved customers*
Demand distribution	*Number of served customers*
Service time	*Sales volume*
Number of servers	*Cost of the service running*
Timetable	*Occupation rate of the server statistics*
Number of service working hours	*Waiting time in a queue statistics*
	Queue size statistics

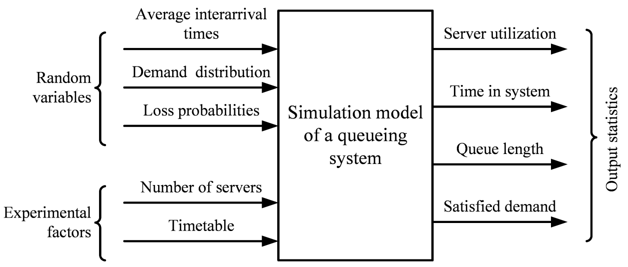
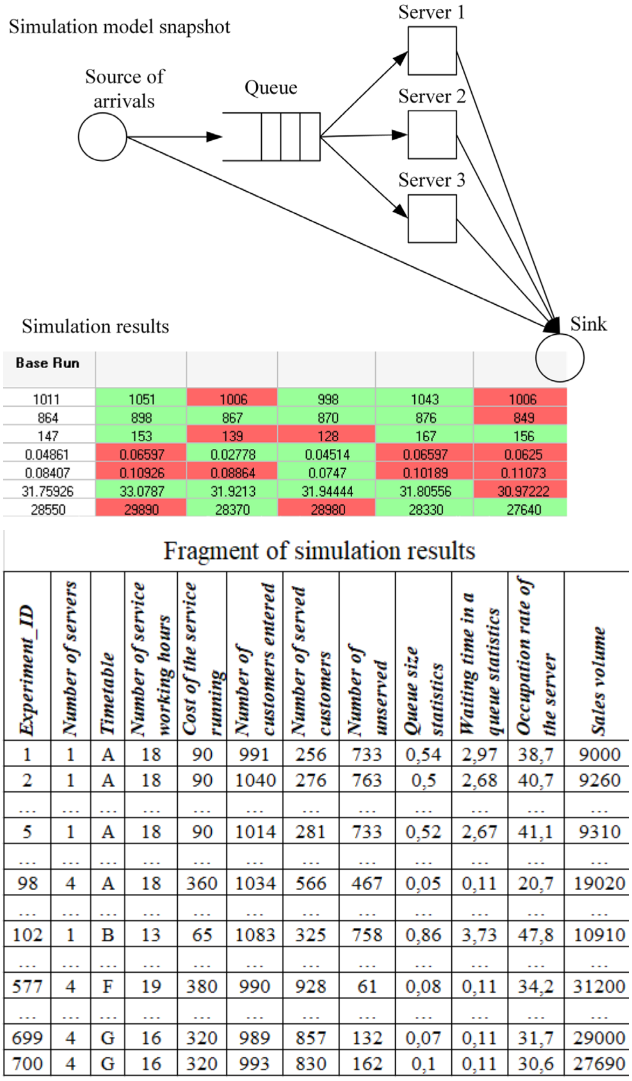


Fig. 3. Conceptual model for the case study.



Base Run					
1011	1051	1006	998	1043	1006
864	898	867	870	876	849
147	153	139	128	167	156
0.04861	0.06597	0.02778	0.04514	0.06597	0.0625
0.08407	0.10926	0.08864	0.0747	0.10189	0.11073
31.75926	33.0787	31.9213	31.94444	31.80556	30.97222
28550	29890	28370	28980	28330	27640

Fragment of simulation results

Experiment_ID	*Number of servers*	*Timetable*	*Number of service working hours*	*Cost of the service running*	*Number of customers entered*	*Number of served customers*	*Number of unserved*	*Queue size statistics*	*Waiting time in a queue statistics*	*Occupation rate of the server*	*Sales volume*
1	1	A	18	90	991	256	733	0,54	2,97	38,7	9000
2	1	A	18	90	1040	276	763	0,5	2,68	40,7	9260
…	…	…	…	…	…	…	…	…	…	…	…
5	1	A	18	90	1014	281	733	0,52	2,67	41,1	9310
…	…	…	…	…	…	…	…	…	…	…	…
98	4	A	18	360	1034	566	467	0,05	0,11	20,7	19020
…	…	…	…	…	…	…	…	…	…	…	…
102	1	B	13	65	1083	325	758	0,86	3,73	47,8	10910
…	…	…	…	…	…	…	…	…	…	…	…
577	4	F	19	380	990	928	61	0,08	0,11	34,2	31200
…	…	…	…	…	…	…	…	…	…	…	…
699	4	G	16	320	989	857	132	0,07	0,11	31,7	29000
700	4	G	16	320	993	830	162	0,1	0,11	30,6	27690

Fig. 4. The simulation model layout, a fragment of simulation report and a fragment of results summary used for data mining.

The general scheme of knowledge discovery consists of three steps:

- Correlation analysis of a relationship between input and output variables.
- Clustering algorithm application against target output variables.
- Results visualization with box charts, histograms and radar charts.

Implementation of these steps has provided the following results.

G. Results of Correlation Analysis

Two output variables with a strong negative correlation *Number of unserved customers* and *Sales volume* have been defined as the target variables. Another important output variable *Cost of the service running* that is correlated with the first two has also been considered. The summary of correlation analysis is provided in Fig. 5.

H. Clustering

The *k*-means clustering method has been implemented for results partition into clusters. In the beginning, the best number of clusters has been determined, which is equal to 4, then the clustering has been carried out, and, finally, cluster ranking has been performed. Clustering has been carried out for three output variables; therefore, the results can be visualised in three-dimensional graphics. *Plotly for R* library has been used for this purpose. The visualization of obtained clusters is provided in Fig. 6. One can see that green and yellow clusters are close in terms of *Number of unserved customers* and *Sales volume*, but the yellow cluster is better in terms of *Cost of the service running*. Both red and orange clusters are worse in terms of all these indicators.

I. Results Visualization

The visualization goal of clustering results is to support the detection of input variable values that are specific for a particular cluster. To detect these values, the box diagrams of two numeric input variables have been created.

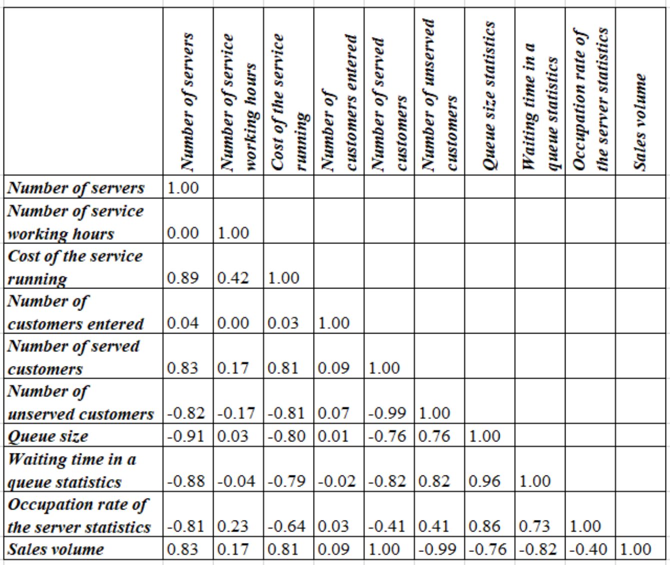
The box diagram of a single input variable *Number of servers* is provided in Fig. 7. Box diagram represents the clusters according to the number of servers. Green cluster is more expensive and requires from three to four servers. With the same number of servers in orange and yellow clusters, the yellow one gives significantly better results for the target output variables. Thus, we can conclude that the yellow cluster is the target one that means the best one in this case.

The next stage of data visualization is carried out through input variable value distribution analysis in clusters. As a result, the value histograms have been created for making a conclusion about the most efficient working timetable. An example of input variable *Timetable* histogram for the target yellow cluster is shown in Fig. 8.

The target cluster histogram of the non-numeric parameter – schedule type – provides the selection of the best parameter value, i.e., B schedule.

The results of this stage may be summarised in an aggregate table and provide a dominant value of this input variable in an appropriate cluster. By using a histogram, it is possible to exclude one of the timetables from further analysis as an inefficient one.

Next, the radar diagrams have been created for all the clusters showing average values of input and output variables. This visualization makes it possible to analyse several variable averages at a time. The radar diagram of one experiment with normalized into the range [0,1] values of input and output variables is shown in Fig. 9. Analysing the radar diagram, it is possible to make sure that one of the clusters is a target cluster.



	Number of servers	*Number of service working hours*	*Cost of the service running*	*Number of customers entered*	*Number of served customers*	*Number of unserved customers*	*Queue size statistics*	*Waiting time in a queue statistics*	*Occupation rate of the server statistics*	*Sales volume*
Number of servers	1.00									
Number of service working hours	0.00	1.00								
Cost of the service running	0.89	0.42	1.00							
Number of customers entered	0.04	0.00	0.03	1.00						
Number of served customers	0.83	0.17	0.81	0.09	1.00					
Number of unserved customers	-0.82	-0.17	-0.81	0.07	-0.99	1.00				
Queue size	-0.91	0.03	-0.80	0.01	-0.76	0.76	1.00			
Waiting time in a queue statistics	-0.88	-0.04	-0.79	-0.02	-0.82	0.82	0.96	1.00		
Occupation rate of the server statistics	-0.81	0.23	-0.64	0.03	-0.41	0.41	0.86	0.73	1.00	
Sales volume	0.83	0.17	0.81	0.09	1.00	-0.99	-0.76	-0.82	-0.40	1.00

Fig. 5. The summary of correlation analysis.

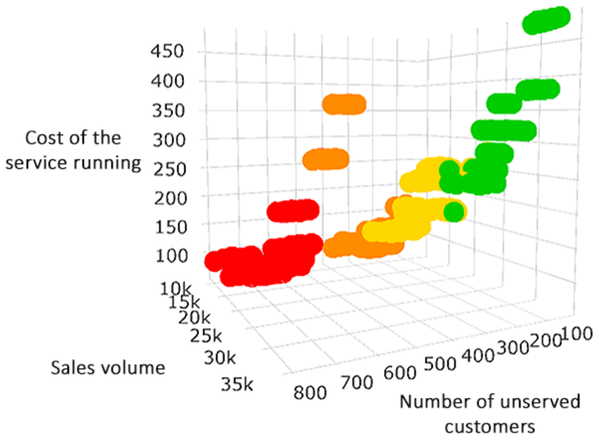


Fig. 6. The results of cluster analysis.

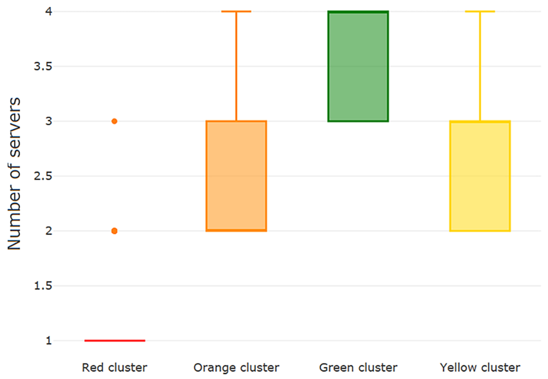


Fig. 7. The box diagram of a single input variable.

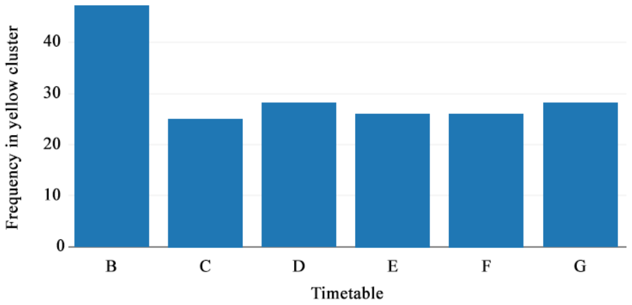


Fig. 8. The histograms of a single input variable values in the target cluster.

J. Decision Rules

Based on the experimental results, the decision principles for the case study problem have been formulated for detecting the input variable values for experimental factors that are specific for each cluster and in particular for the best cluster. The values for the input variables that are experimental factors of the best cluster are shown in Table II.

Thus, the main goal of the case study – the best performance of a queueing system – has been achieved.

V. Conclusion

A variety of data mining techniques, including correlation analysis, clustering and several visualization mechanisms of results, have been applied during knowledge discovery.

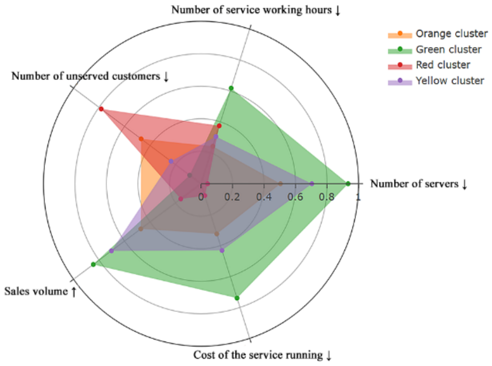


Fig. 9. The radar diagram of input and output variables of four clusters in a single experiment.

TABLE II

The Obtained Values for the Input Variables of the Best Cluster

Input variables: Experimental factors	Values obtained for the best cluster
Number of servers	[2–3]
Number of service working hours	[13–19]
Timetable	B

Experiments that correspond to the defined combination of data mining techniques have been designed. The results of the experiments have supported the identification of the relationships between the simulation model input and output variables and definition of target output variables. All data records have been arranged into groups according to the values of target output variables and the best group has been chosen. According to experiment results, the values of the input variables that are typical of each created group, particularly of the best group, have been defined. Knowledge and decision rules have been obtained from simulation results. Finally, the following conclusions have been drawn:

- The target output variable selection, which is carried out by clustering, is a decisive phase. This selection determines the results of a further analysis, as well as the speed and efficiency of the procedure in general.
- If there are no strictly defined target values of the output variables, the most relevant technique may be clustering.

The simulation experiments performed in the research have obtained useful knowledge from simulation and confirmed the initial hypothesis about the suitability of the proposed scheme for efficient simulation results analysis, knowledge discovery and decision formulation. To conclude, the developed scheme is applicable to problem-solving in queuing systems as well as in other simulation-based analysis projects.

References

[1] M. Pidd, *Systems Modelling: Theory and Practice*. Chichester: John Wiley & Sons Ltd, 2004, 192 p.

[2] A. M. Law, and W. Kelton, *Simulation modeling and analysis,* 3rd ed. Mc Graw Hill Higher Education, 2000.

[3] J. Banks, and J. S. Carson, B. L. Nelson, & D. Nicol, *Discrete-event System Simulation*, 5th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[4] J. Merkurjevs, J. Pečerska, and J. Tolujevs, “Simulation-Based Analysis of Logistic Systems,” *Humanities and Social Sciences. Latvia*, vol. 4, no. 57, pp. 27–48, 2008.

[5] T. F. Brady, and R. A. Bowden, “The effectiveness of generic optimization routines in computer simulation languages,” *Proceedings of the 2001 Industrial Engineering Research Conference*, 2001.

[6] T. F. Brady, E. Yelling, “Simulation data mining: A new form of computer simulation output,” *Proceedings of the 2005 Winter Simulation Conference*, 2005. https://doi.org/10.1109/WSC.2005.1574262

[7] N. Feldkamp, S. Bergmann, S. Strassburger, “Knowledge Discovery in Manufacturing Simulations,” *Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, 2015. https://doi.org/10.1145/2769458.2769468

[8] N. Feldkamp, S. Bergmann, S. Strassburger, T. Schulze, “Knowledge discovery in simulation data: A case study of a gold mining facility,” *Proceedings of the 2016 Winter Simulation Conference*, 2016. https://doi.org/10.1109/WSC.2016.7822210

[9] D. Kibira, Q. Hatim, S. Kumara, et al. “Integrating data analytics and simulation methods to support manufacturing decision making,” *Proceedings of the 2015 Winter Simulation Conference*, 2015. https://doi.org/10.1109/WSC.2015.7408324

[10] M. K. Painter et al., "Using simulation, data mining, and knowledge discovery techniques for optimized aircraft engine fleet management," *Proceedings of the 2006 Winter Simulation Conference*, 2006. https://doi.org/10.1109/WSC.2006.323221

[11] M. Dunham, *Data Mining: Introductory and Advanced Topics*. Pearson Education, Inc., 2003, p. 315.

[12] C. G. Cassandras, and S. Lafortune, *Introduction to Discrete Event Systems, Second edition*., Boston, MA, USA: Springer, 2008. https://doi.org/10.1007/978-0-387-68612-7

[13] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques, Second Edition*, San Francisco, CA, USA: Elsevier Inc., 2006.

[14] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*, San Francisco, CA, USA: Elsevier Inc., 2005.

[15] Y. Zhao, *R and Data Mining: Examples and Case Studies, First edition*, Elsevier Inc., 2013.

[16] I. Sitova, and J. Pecherska, "A concept of simulation-based SC performance analysis using SCOR metrics," *Information Technology and Management Science*. vol. 20, no. 1, pp. 85–90, 2017. https://doi.org/10.1515/itms-2017-0015

[17] S. Robinson, "A tutorial on conceptual modeling for simulation," *Proceedings of the 2015 Winter Simulation Conference*, 2015. https://doi.org/10.1109/WSC.2015.7408298

Jeļena Pečerska is a Doctor of Information Technology in the field of system analysis, modelling and development. She has been working for Riga Technical University since 1979. At the moment, she is an Associate Professor at the Department of Modelling and Simulation of Riga Technical University. Professional interests include methodology of discrete-event simulation, combined simulation, supply chain modelling, practical applications of discrete-event simulation and discrete-event simulation in education. She is a member of the Latvian Simulation Society.
Email: jelena.pecerska@rtu.lv
ORCID iD: https://orcid.org/0000-0002-7779-6305

Irīna Šitova is a Doctoral student at the Information Technology Institute, Riga Technical University (RTU). She received her *M. sc. ing.* degree in Information Technology from RTU in 2018. Currently, she is a Scientific Assistant at the Department of Modelling and Simulation of RTU. Her interests include discrete-event simulation, decision making, data mining and machine learning.
Email: irina.sitova@rtu.lv
ORCID iD: https://orcid.org/0000-0002-3035-6468