RESEARCH IN LOGISTICS & PRODUCTION

A FRAMEWORK FOR ADDITIONAL SERVER ACTIVATION

Gordana Savić*, Dragana Makajić-Nikolić*, Mirko Vujošević*

*University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia,
E-mail: goca@fon.bg.ac.rs,
E-mail: gis@fon.bg.ac.rs,
E-mail: mirkov@fon.bg.ac.rs

Abstract This paper assesses the performance of the queuing system with several fixed and additional multitask servers. The additional server serves customers in the case of necessity to avoid congestion and decrease the number of impatient customers. An additional server will be activated whenever the length of queue becomes greater than or equal to user defined size i.e. threshold value. Also, this server will be switched off when the length of queue falls below user defined size, but ongoing service has to be finished.The main issue is to decide when and how long the additional server would work with customers. This paper proposes a methodology to choose the best activation rule for the work of an additional server, based on the performance evaluation of queuing system. The Petri nets (PN) simulation is used to calculate the performance measures. Relative efficiency evaluation of different activation rules is carried out by data envelopment analysis (DEA). The methodology is applied on real-life data collected from a postal office.

Paper type: Research Paper

Published online: 29 October 2012
Vol. 2, No. 4, pp. 387-397

ISSN 2083-4942 (Print)
ISSN 2083-4950 (Online)

Keywords: *Queue, Additional server, Activation rule, Petri nets, Data envelopment analysis*

1. INTRODUCTION

This paper is motivated by a practical system of a post office, which offers different types of services. Work with customers is organized as a single line queuing system, with c fixed servers and one additional server-counter ($d = 1$) The structure of the system can be described as: the single line, multi-server, queuing system with an additional server. An activation rule assumes that the additional server is activated whenever the length of queue exceeds a threshold value.

One of the main features of the observed system is the stochasticity. The arrival of customers and requests for a particular service are stochastic. The customers are served in the order of their arrival; that is, the first-come, first-served discipline. The service performing time depends on the service type and it is also stochastic. A percentage of customers are treated as impatient and they will abandon the system immediately upon arrival depending entirely on the length of queue. It is assumed that there are no retrials.

A comparison of system efficiency for different threshold values should be done. The PN is used for queuing system simulation and calculation of performance measures. Afterwards, relative efficiency for different thresholds is evaluated by DEA.

The queuing systems with characteristics similar to those described above have been analyzed in several studies. An optimal threshold level was numerically computed for multi-server retrial queues with variable numbers of servers in (Artlejo, Orlovsky, & Dudin, 2005). Kaboudan (1998) developed a queuing simulation algorithm for adjusting the number of servers in a single line, multi-server system. Authors in (Zhang & Tian, 2004) studied the Markovian queuing system, where each server performs services as a primary or taking vacation as a secondary job. Jain, Sharma and Shekhar (2005) analysed a finite, queue-dependent, heterogeneous, multi-processor service system with limited service capacity in which processors were shared by more than one job. The authors in (Shin & Choo, 2009) developed the retrial model with finite capacity of service facility by assigning specific values to the probabilities on which customers join or abandon the system or retry to be served.

A wide class of queuing systems have been modelled and simulated by PN. A multi-server and multi-queue network for distributed systems were studied in (Shan, Lin, & Yang, 2002). GSPN was used in (Gharbi & Ioualalen, 2010) to obtain exact performance measures of finite-source, multi-server queuing systems with different vacation policies..

DEA was applied for evaluation of several operating scenarios in (Srdjevic, Medeiros, & Porto, 2005). DEA was also employed in evaluating of relative efficiency of eleven global climate policy scenarios in (Bosetti & Buchner, 2009). As a ranking method, DEA was also employed for a selection of dispatching rules with respect to several performance criteria in (Braglia & Petroni, 1999).

The rest of the paper is organized as follows. The fundamentals of Petri nets and DEA are given in Section 2. A description of proposed evaluation methodology given in Section 3 is followed by an example of an application in a post office (Section 4). Section 5 discusses methodology and offers final conclusions.

2. BACKGROUND

Petri nets

A Petri net is a particular kind of a directed graph, together with an initial state called initial marking. The underlying graph of a Petri net is a directed, weighted, bipartite graph consisting of two kinds of nodes, called places and transitions. Coloured Petri nets (CPNs) is a discrete-event modelling language combining the capabilities of Petri nets with the capabilities of a high-level programming language (Jensen & Kristensen, 2009). CPN is executable, allowing the flow of tokens around the net to be visualised. This can illustrate the flows of data within the same model. For a practical use, CPN is represented by a graphic form which allows visualization of system dynamics (flows of data and control). The graphical form comprises two parts: a *Graph* which represents the net elements graphically and carries textual inscriptions; and a *Declaration*, defining all the types, variables, constants and functions that are used to annotate the Graph part.

The following CPN extensions are used in order to model a queuing system: hierarchical CPN and Generalized Stochastic CPN (GSPN). GSPN are timed CPN in which some transitions are timed and stochastic while others are immediate.

Data envelopment analysis

DEA has been widely used for evaluating the relative performance of similar decision making units (DMUs) with multiple inputs and outputs. The original DEA model was given by authors, Charnes, Cooper and Rhodes (1978), who tried to generalize single-input to single-output ratio definition of efficiency to ratio of sum of weighted outputs to sum of weighted inputs. Suppose that DMU_j ($j = 1,\ldots,n$), within set of n units, uses inputs x_{ij} ($i = 1,\ldots,m$) to produce outputs y_{rj} ($r = 1,\ldots,s$), absolute efficiency measure (Podinovski, 1999) is as follows:

$$E_j = \sum_{r=1}^{s} u_r y_{rj} \Big/ \sum_{i=1}^{m} v_i x_{ij}$$

where v_i ($i = 1,\ldots,m$) are input multipliers and u_r ($r = 1,\ldots,s$) are output multipliers (weights). The LP weighted form of the basic DEA CCR model is as follows:

$$(\max)\, h_k = \sum_{r=1}^{s} u_r y_{rk}$$

$$\sum_{i=1}^{m} v_i x_{ij} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0,\ j = 1,\ldots,n$$

$$u_r \ge \varepsilon;\ v_i \ge \varepsilon;\ r = 1,2,\ldots,s,\ i = 1,2,\ldots,m$$

As a solution of basic CCR DEA models, efficiency score h_k is 1 for all efficient units and lower than 1 for all inefficient units. A lot of methodologies are designed to allow ranking in DEA context. For example, Andersen and Petersen introduced a super-efficiency measuring model (Andersen & Petersen, 1993) by excluding inputs and outputs of DMU_k from constraints (4). Regardless the DEA model used, multipliers are determined objectively and distinctively for each DMU. But, a well-known idea take into account a priori weight restriction is assurance region method (AR) introduced by (Thompson, Singleton, Trall, & Smith, 1986):

$$L_i \le v_i / v_{i+1} \le U_i$$

A ranking based on absolute efficiency values (1), DEA model (2)-(5) and MCDM Weighted Sum Model (WSM) to verify results are used. Procedure will be described in details in the last phase of methodology.

3. EVALUATION METHODOLOGY

This section proposes a methodology for a simulation and an efficiency evaluation of a single queue, multi-server and multi-service system on different threshold values for a given activation rule. The methodology framework consists of three straightforward phases (Fig. 1).

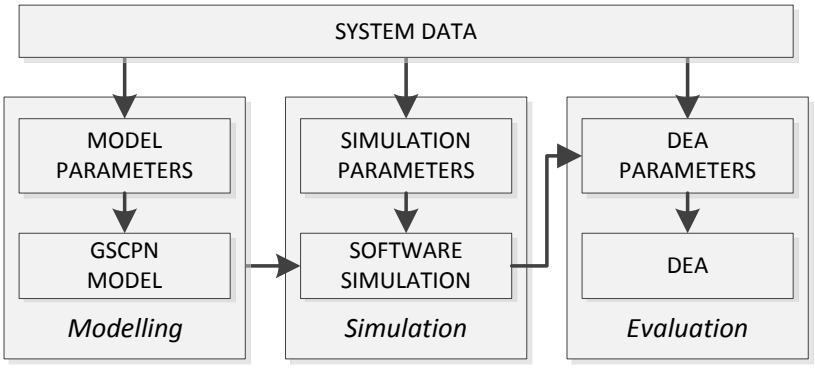

Fig. 1 Evaluation methodology

The first step of each phase is a specification and measurement of parameters. Some of the parameters are predefined as system data, and others are obtained as performance measures from the previous phase. A list of performance measures will be suggested in the following description.

Phase 1. ***Queuing system modelling***

GSPN is used for modelling of the queuing system. General net structure is designed for the queuing system characterized with the following features: single line queuing system with *c* fixed servers and an additional one, first-come, first-served queuing discipline, stochastic inter-arrival time, stochastic occurrence of services types, stochastic serving time, stochastic customer impatience related to length of queue and activation rule set as length of queue.

The arrivals and services are modelled by two sub-pages of hierarchical Generalized Stochastic Coloured Petri net (GSCPN) with a queuing place as their connection. Queuing system parameters are defined in this phase of system modelling and in the next phase of simulation. Some of queuing system features are represented as PN model parameters (fixed and additional servers, service, queue, customer arrivals, impatient customers) while the others are simulation parameters (system capacity, inter-arrival time distribution, number of service types, service type frequency, service time distribution and threshold value) which are controllable parameters.

Phase 2. Simulation

The system performance measures are calculated as average of mean values collected from independent replications of simulation. As an output of the simulation, different performance measures can be obtained. One set of them is standard queuing system performance measures such as: length of queue, waiting time, number of customers in system and time spent in system. The other set of performance measures is specific for the observed queuing system. Some of these measures are: number of served customers, total work hours and overtime, fixed servers occupancy and number of additional server activations.

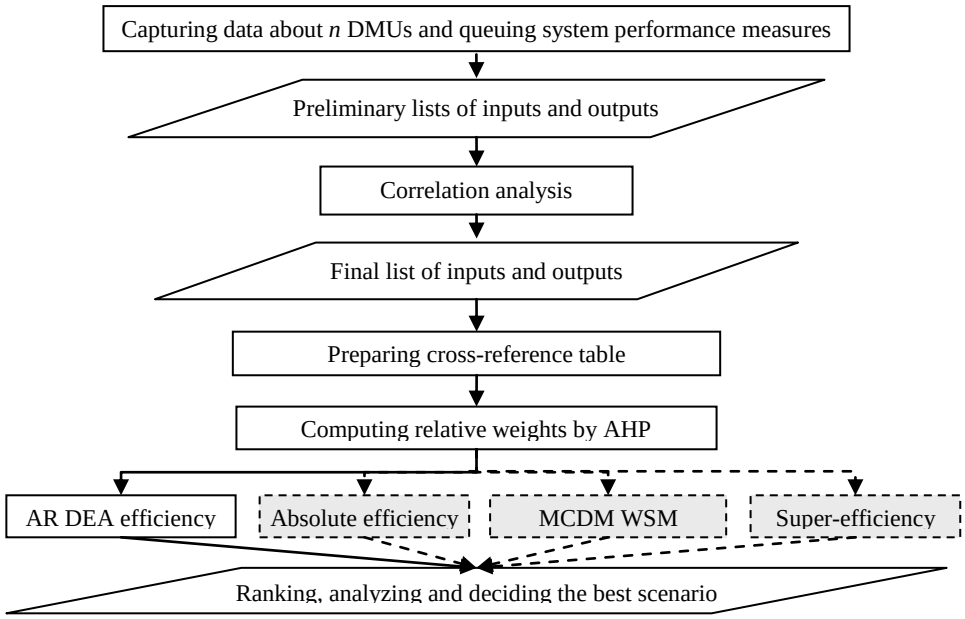

Fig. 2 Flowchart of scenarios evaluation and results verification

Phase 3. Evaluation and ranking of scenarios

The third phase in the methodology is DEA performance evaluation and ranking of scenarios based on performance measures gained form the previous phase.

Flowchart of evaluation and ranking process is shown in Fig. 2. A parameters definition starts with choosing scenarios as DMUs and selecting DEA inputs and outputs. The first step is capturing simulated performance measures on *n* DMUs. In the second step, decision maker should decide whether the particular performance is a candidate to be DEA input or output. With the aim of checking data consistency and justifying division on inputs and outputs sets, correlation analysis should be done in the third step. The next recommended step is the revision and finalization of the list of selected inputs and outputs. The final step in preparing parameters is making cross-referencing data table for scenarios and the criteria.

The comparison of scenarios rank according to efficiency ratios is assessed by several DEA and MCDM models according the flowchart in Fig. 2. One branch of evaluation is the ranking when weight restrictions, obtained by AHP, are included into the model. The other way is to calculate weighted outputs and inputs and finally, the absolute efficiency using the Eqs. (1). Third, super-efficiency ratios, obtained using Andersen-Petersen's model allows direct ranking. In additionWeighted Sum Method (WSM) is used for results verification. These verification methods are optional and are shown in shadowed rectangles on the flowchart Fig. 2.

4. CASE STUDY

The system parameters used in this case study are calculated based on 10 days observation at one Serbian post office, where there is a congestion problem.

Phase 1: Queuing system modelling

In the post office two counters work with customers during the whole working day. Another counter, which primarily works as a back office, can also serve customers whenever the number of customers in a queue exceeds threshold value. Also, the additional counter will be back to its primary job whenever the queue is empty. The GSCPN model of the post office is implemented as described in section 3 and designed in CPN tools (Jensen, Kristensen, & Wells, 2007).

Phase 2: Simulation

The model designed in CPN tools is suitable for simulation of queuing system with different values of parameters. It is assessed that inter-arrival time fits to exponential distribution with expected value of 125 seconds, performed on the real data in SPSS. Most of the customers have asked just for one of the twelve types of postal services. SPSS tool is also used to confirm that the service time fits exponential distribution for each particular service type. Frequencies of occurrence of different service types and expected service time are given in Table 1. Since the expected service time of all service types is 277.58 seconds, occupation rate per fixed counter is 1.11. Thereafter, it is obvious that such a system can achieve unsteady state and temporary activation of an additional counter is needed.

Table 1 Service type parameters

Service type	1	2	3	4	5	6	7	8	9	10	11	12
Frequencies	.243	.152	.072	.044	.047	.072	.088	.099	.075	.050	.033	.025
Time (seconds)	139.7	230.6	224.9	243.1	559.0	585.2	747.3	89.9	188.8	213.4	226.0	250.6

Besides these, two more parameters are set up. The first is abandoning frequencies depending on the length of a queue (Table 2) and the second is the system capacity of 27 customers in total.

Table 2 Abandoning probability

Length of queue	0-3	4-6	7-9	10-12	12-15	16-18	19-21	22-24	>24
Frequencies	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0

The threshold value is chosen as controllable variable in observed case and each threshold value is set up as one scenario. The thirteen scenarios are created by activating an additional counter whenever the length of queue reaches predefined threshold values from 3 to 15. For each scenario, 50 independent simulations are replicated. Obtained performance average measures are used in the evaluation phase.

Phase 3. ***Evaluation and ranking of scenarios***

Following the flowchart in Fig. 6, data from simulation are captured. The threshold values are considered as thirteen DMUs and performance measures are taken as criteria. The final list of criteria consists of 2 inputs and 3 outputs given in table 3: The cross-referencing table, made for these five criteria and thirteen DMUs, is used as a DEA data table.

Table 3 Criteria and weights

Criteria	Set 1	Set 2
Length of queue (I)	0.308	0.333
% of impatient customers (I)	0.192	0.167
# of additional counter activations (O)	0.076	0.200
Additional counter availability (O)	0.315	0.200
Fixed counters availability (O)	0.109	0.100
Inconsistency level	0.1	0.0

In the next step, Expert Choice software is employed for the calculation to gain two different sets of criteria weights given in Table 3. The Set 1 follows the business policy of prioritizing serving of clients, even if that leads to increasing number of additional counter activations. The Set 2 corresponds to the business policy of cutting costs or the time consumed during the activation and deactivation of an additional counter.

Table 4 DEA efficiency scores

DMUs (Scenarios)	Threshold values (length of queue)	Super efficiency	AR DEA	
			Weights Set 1	Weights Set 2
DMU1	**3**	**1.118**	**1.000**	**1.000**
DMU2	**4**	**1.107**	0.913	0.928
DMU3	**5**	0.998	0.951	0.984
DMU4	**6**	**1.031**	0.933	0.978
DMU5	**7**	0.973	0.898	0.951
DMU6	**8**	0.882	0.797	0.871
DMU7	**9**	0.864	0.761	0.853
DMU8	**10**	0.938	0.806	0.920

DMU9	**11**	0.952	0.794	0.935
DMU10	**12**	0.882	0.720	0.872
DMU11	**13**	0.917	0.717	0.913
DMU12	**14**	**1.045**	0.741	**1.000**
DMU13	**15**	**1.067**	0.689	0.974

Furthermore, ratios between two input weights or two output weights become the lower bounds in Eqs. (6) and the upper bounds are set to 10. Two groups of DEA models (1)-(6) with restrictions corresponding to Set1 and Set2 are solved by DEA-solver software (Cooper, Seiford, & Tone, 2006). The relative efficiency scores, for both business policies aforementioned, are presented in Table 4 as well as the scores gained by solving super-efficiency DEA model.

According to super-efficiency scores, five out of thirteen scenarios are found out as efficient and reach values greater or equal to 1. This result is a consequence of the total weights flexibility in DEA. The AR DEA model with Set 1 identifies DMU1 as distinctly the top ranked threshold value. In the case of weights Set 2, the obtained results indicate that DMU1 and DMU12 are relatively efficient. The smaller length of queue and consequently, shorter expected waiting time and smaller percentage of impatient customers, affected greatly efficiency regardless the importance assigned to the criteria. As a result, Scenario DMU1 (threshold value of 3) remains efficient even when the number of additional counter activations has the same importance as its availability. On the other hand, DMU12 (threshold value of 14) is efficient as a consequence of the small number of additional counter activations (1.68 average).

5. DISCUSSION AND CONCLUSION

Unsteady state of the queuing system has been examining in this paper. The unsteady state is related to congestion, as a consequence of customer arrivals and service rate imbalance. Another negative consequence is the impatient customers. The observed system avoids a congestion problem by temporary activation of an additional server depending on the length of a queue. The methodology for better allocation of an additional server serving time is based on PN simulation and DEA efficiency evaluation as proposed. The results obtained throughout the procedure could be used as a decision making support to the management of a dynamic queuing systems. A slight modification in the PN model would enable application of the methodology proposed in this paper in different queuing systems, such as banks, mega markets, passport and customs controls, road toll or IT networks. The rules could be set, for example, as a requested type of services or inter-arrival peak time. The number of queues, queuing discipline, customer's impatience and retrial behaviour could be changed. The third phase of methodology can be applied

in interactive procedures, by changing the aspiration level to queuing system performance measures in order to come to the solution.

ACKNOWLEDGEMENT

This research was partially supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, Projects: TR35045 and III44007.

REFERENCES

Andersen, P., & Petersen, N. (1993). A procedure for ranking efficient units in data envelopment analysis. Management Science, Vol. 39(10), pp. 1261-1264.

Artlejo, J.R., Orlovsky, D.S., & Dudin, A.N. (2005). Multi-server retrial model with variable number of active servers. Computers&Industrial Engineering, Vol. 48, pp. 273-288.

Bosetti, V., & Buchner, B. (2009). Data envelopment analysis of different climate policy scenarios. Ecological Economics Vol. 68, pp. 1340–1354.

Braglia, M., & Petroni, A. (1999). Data envelopment analysis for dispatching rule selection. Production Planning & Control, Vol. 10, No. 5., pp. 454-461.

Charnes, A., Cooper, W. W., & Rhodes, E. L. (1978). Measuring the efficiency of decision making units. European Journal of Operational Research, Vol. 2(6), pp. 429-444.

Cooper, W. W., Seiford, L. M., & Tone, K. (2006). Introduction to data envelopment analysis and its uses with DEA-solver software and references. Springer.

Gharbi, N., & Ioualalen, M. (2010). Numerical investigation of finite-source multiserver systems with different vacation policies. Journal of Computational and Applied Mathematics Vol. 234, pp. 625-635.

Jain, M., Sharma, G. C., & Shekhar, C. (2005). Processor-shared service systems with queue-dependent processors . Computers & Operations Research Vol. 32, pp. 629–645.

Jensen, K., & Kristensen, L. M. (2009). Coloured Petri nets. Modelling and validation of concurrent systems. Springer.

Jensen, K., Kristensen, L. M., & Wells, L. (2007). Coloured Petri nets and CPN Tools for modeling and validation of concurrent systems. International Journal on Software Tools for Technology Transfer (STTT), Vol. 9, Numbers 3-4, pp. 213-254.

Kaboudan, M. A. (1998). A dynamic-server queuing simulation. Computers Ops Res., Vol. 25(6), 431-439.

Lin, H. T. (2010). Personnel selection using analytic network process and fuzzy data envelopment analysis approaches. Computers & Industrial Engineering Vol. 59, pp. 937-944.

Podinovski, V. V. (1999). Side effects of absolute weight bounds in DEA models. EJOR Vol. 115, pp. 583-595.

Roll, Y., Cook, W. D., & Golany, B. (1991). Controlling factor weights in data envelopment analysis. IIE Transactions Vol. 23, pp. 2-9.

Shan, Z., Lin, C.,Yang, Y. (2002). A multiserver multiqueue network: modeling and performance analysis. Journal of University of Science and Technology Beijing, pp. 389-395.

Shin, Y. W., & Choo, T. S. (2009). M/M/s queue with impatient customers and retrials. Applied Mathematical Modelling Vol. 33, pp. 2596–2606.

Srdjevic, B., Medeiros, Y. D., & Porto, R. L. (2005). Data envelopment analysis of reservoir system performance. Computers & Operations Research Vol. 32, pp. 3209–3226.

Thompson, R. G., Singleton, F. D., Trall, R. M., & Smith, B. A. (1986). Comparative site evalutaion for locating a high-energy physics lab in Texas. Interfaces Vol.16, pp. 34-49.

Zhang, Z. G., & Tian, N. (2004). An analysis of queueing systems with multi-task servers. EJOR Vol. 156, pp. 375-389.

BIOGRAPHICAL NOTES

Gordana Savić completed her Ph.D. in Operations Research at the University of Belgrade, Faculty of Organizational Sciences in 2012. She works as professor assistant at the Faculty of Organizational Sciences at the University of Belgrade, where she lectures courses in the Operations Research field. Her research interests are related to Mathematical Modeling, Optimization methods and DEA theory and application. She is the author or co-author of over 40 papers and co-author of two books in the OR area. Her papers have been published in journals including European Journal of Operational Research and Scientometrics.

Dragana Makajić-Nikolić completed her Ph.D. in Operations Research at the University of Belgrade, Faculty of Organizational Sciences in 2012. She works as professor assistant at the Faculty of Organizational Sciences at the University of Belgrade, where she lectures courses in the Operations Research field. Her research interests are related to Mathematical Modeling, Optimization methods and Risk Analysis. She is the author or co-author of over 40 papers and two books in the OR area. Her papers have been published in journals including Energy Policy.

Mirko Vujošević graduated in electrical engineering at Belgrade University where he finished his postgraduate studies and earned his doctorate. From 1976 to 1995 he was with Mihailo Pupin Institute, Belgrade, and now he is full professor at the Faculty of Organizational Sciences, Belgrade University. He published more than 180 professional papers on different topics of operational research, reliability, maintenance, inventory control and applied mathematics. He is author and co-author of two monographs, six textbooks, and several chapters in monographies.