European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Performance improvement of a service system via stocking perishable preliminary services

Gabi Hanukov[a], Tal Avinadav[a,*], Tatyana Chernonog[a], Uri Yechiali[b]

[a] Department of Management, Bar-Ilan University, Ramat Gan 5290002, Israel
[b] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

ARTICLE INFO

Article history:
Received 12 February 2018
Accepted 12 October 2018
Available online xxx

Keywords:
Queueing
Preliminary services
Perishable products
Inventory
Food industry

ABSTRACT

The typical fast food service system can be conceptualized as a queueing system of customers combined with an inventory of perishable products. A potentially effective means of improving the efficiency of such systems is to simultaneously apply time management policies and inventory management techniques. We propose such an approach, based on a combined queueing and inventory model, in which each customer's service consists of two independent stages. The first stage is generic and can be performed even in the absence of customers, whereas the second requires the customer to be present. When the system is empty of customers, the server produces an inventory of first-stage services ('preliminary services'; PSs) and subsequently uses it to reduce future customers' overall service and sojourn times. Inventoried PSs deteriorate while in storage, creating spoilage costs. We formulate and analyze this queueing-inventory system and derive its steady-state probabilities using matrix geometric methods. We show that the system's stability is unaffected by the production rate of PSs. We subsequently carry out an economic analysis to determine the optimal PS capacity and optimal level of investment in preservation technologies.

1. Introduction

There are service systems in which the service is composed of two stages, where the first stage can be carried out before the arrival of customers and its products stocked until they are required or until they spoil in storage. Notable examples of such systems are observed in the fast food industry, an industry in which mass-produced food is prepared and served quickly in dining establishments (e.g., McDonalds, Starbucks, KFC, as well as food trucks). In fast food restaurants, it is common for food products to be processed in advance and stored (e.g., pre-cooked hamburger patties, chopped vegetables, etc.), such that when the customer arrives, the server only needs to take a few steps to complete the food service (e.g., heating the patty, adding sauce, arranging the dish, etc.; http://science.howstuffworks.com/innovation/edible-innovations/fast-food.htm). The revenue of the fast food industry amounts to hundreds of billions of dollars per year (http://www.restaurant.org/News-Research/Research/Facts-at-a-Glance), such that even a small improvement in the efficiency of fast food services has the potential to generate vast savings. The current study explores means of obtaining such an efficiency improvement in a fast food service system that incorporates a two-stage production process.

In general, service in a fast food restaurant can be conceptualized as a queueing system of customers combined with an inventory of perishable products. Accordingly, a potentially effective approach to improve the efficiency of fast food service would be to target both time management—that is, management of customers' waiting time and servers' idle time—and inventory management, in a "pincer movement". Strategies for reducing customer waiting times in queueing systems include using faster servers (e.g., Guo & Zhang, 2013, Hwang, Gao, & Jang, 2010), serving customers in batches (Zacharias & Pinedo, 2017) rather than one at a time (Shi & Lian, 2016), or adding servers to the system. However, these strategies are likely to increase servers' idle time. Numerous studies have proposed eliminating some of the waste associated with servers' idle time by utilizing idle servers for ancillary duties, so-called "vacations" (see, e.g., Boxma, Schlegel, & Yechiali, 2002; Levy & Yechiali, 1975, 1976; Mytalas & Zazanis, 2015; Rosenberg & Yechiali, 1993; Yang & Wu, 2015; Yechiali, 2004; and Guha, Goswami, & Banik, 2016). However, this approach might negatively affect customers' waiting time, e.g., if "vacation" tasks take a substantial amount of time to complete and prevent servers from returning to their primary duties upon a customer's arrival. In light of these considerations, it is a serious challenge to reduce customers' waiting time and servers' idle time simultaneously in service systems.

* Corresponding author.
E-mail addresses: german.khanukov@live.biu.ac.il (G. Hanukov), tal.avinadav@biu.ac.il (T. Avinadav), tatyana.chernonog@biu.ac.il (T. Chernonog), uriy@post.tau.ac.il (U. Yechiali).

https://doi.org/10.1016/j.ejor.2018.10.027
0377-2217/

Fast food providers attempt to meet this challenge by decomposing the food preparation process as elaborated above, preparing and storing food products in advance and completing the service in the presence of the customer. Hanukov, Avinadav, Chernonog, Spiegel, and Yechiali (2017, 2018) have described a model that, to some extent, captures this process: In their model, a server exploits its idle time to produce so-called "preliminary services" (PSs) and store them, with the goal of providing future customers with faster service when they arrive (as opposed to executing the entire service from beginning to end in the presence of the customer). However, their model does not take into account the possibility of product deterioration when considering food industries. Food products, and especially fresh ingredients, might spoil if kept too long in storage prior to customers' arrival. Indeed, according to Gustavsson, Cederberg, Sonesson, van Otterdijk, and Meybeck (2011), waste is a serious problem in food industries: approximately one-third of the food produced worldwide is wasted or lost annually. Proper inventory management of perishable products has the potential to play an important role in reducing losses due to waste.

Inventory management of perishable items is extensively discussed in the literature (see, e.g., Avinadav & Arponen, 2009; Avinadav, Chernonog, Lahav, & Spiegel, 2017; Avinadav, Herbon, & Spiegel, 2013, 2014; Berk & Gürler, 2008; Chao, Gong, Shi, & Zhang, 2015; Chen, Pang, & Pan, 2014; Chernonog and Avinadav 2017; Cooper, 2001; Herbon 2018; Herbon & Khmelnitsky, 2017; Hu, Shum, & Yu, 2015; Li, Yu, & X, 2016; Zhang, Shi, & Chao, 2016). Most studies in the operations management literature consider the deterioration rate of perishable products to be an exogenous variable, which is not subject to control (see, e.g., Hsieh & Dye, 2017; Pahl & Voß, 2014; Wang, Teng, & Lou, 2014). Yet, in practice, the deterioration rate of perishable products can be controlled and reduced via various methods, such as process changes and acquisition of improved preservation technologies. Indeed, according to Dye and Hsieh (2012), many enterprises have studied causes of deterioration and developed preservation technologies to control it and increase their profits. Only a few academic studies have considered the use of preservation technology as an endogenous variable (Dye, 2013; Dye & Hsieh, 2012; Hsu, Wee, & Teng, 2010; Yang, Dye, & Ding, 2015; Zhang, Wei, Zhang, & Tang, 2016). According to these studies, investment in preservation technologies can enable firms to reduce economic losses and improve customer service, and thereby obtain a competitive advantage.

We propose to model the service process in a fast food establishment as a queueing-inventory system, where the queue consists of customers waiting to be served, and the inventory includes food products that the server partially prepares and stores (PSs) during time periods in which there are no customers in the system. This inventory has the potential to reduce the sojourn times of customers, who, instead of waiting for the full service (FS) to be completed from scratch (e.g., waiting for a raw hamburger patty to be cooked and placed into a bun) need to wait only for a "complementary service" (CS) to be carried out on the pre-prepared food product (e.g., waiting for a pre-cooked hamburger patty to be placed into a bun). Our model further assumes that the partially prepared products (PSs) undergo a deterioration process while in storage.

Herein, we formulate and analyze this model and determine its steady-state probabilities. Next, given that the decision variable is the maximum number of items that can be stored in inventory (denoted by n), we define several performance measures for the system (including the number of customers in the system; the number of inventoried PSs in the system; mean sojourn times of customers and of inventoried products) and show how they can be calculated as a function of n. We subsequently assign a cost to each measure and derive the value of n that minimizes the total costs associated with the system. Finally, we carry out a similar analysis for a more complex case in which the decision maker can control products' deterioration rate by investing in preservation technologies.

To the best of our knowledge, a model that combines a queueing system and a perishable-inventory system built into a multi-stage service has not appeared in the literature. Prior works that have considered combinations of queueing systems and inventory-type systems include those of Zhao and Lin (2011), and Adacher and Cassandras (2014). Jeganathan, Reiyas, Padmasekaran, and Lakshmanan (2017) investigated a system consisting of a perishable inventory that uses a two-rate service policy within a finite queueing system under a continuous review (s, Q) ordering policy. Nair, Jacob, and Krishnamoorthy (2015) considered a multi-server Markovian queuing model where each server provides service only to one customer, and the servers are considered as an inventory that will be replenished according to the standard (s, S) policy. Krishnamoorthy, Manikandan, and Lakshmy (2015) considered two control policies, (s,Q) and (s, S), for an $M/M/1$ queueing-inventory system, in which the item is given with probability γ to a customer at his service completion epoch. The authors investigated optimization problems associated with both models. Avşar and Zijm (2014) analyzed base-stock controlled multi-stage production-inventory systems with capacity constraints and gave an approximated solution based on recursive algorithms. Altendorfer and Minner (2015) modeled a production system as an M/M/1 queue with input rates that are dependent on queue length and random customer required lead time, and they developed a heuristic solution for the optimal capacity investment problem. Chebolu-Subramanian and Gaukler (2015) used queueing theory to analyze product contamination in a multi-stage food supply chain with inventory.

Another research stream that is closely related to our work investigates queues in which the service time of each individual customer is composed of multiple phases (stages). Several studies in this stream (Choudhury, 2007, 2008; Choudhury & Deka, 2012; Choudhury & Madan, 2005; Choudhury, Tadj, & Paul, 2007) have investigated M/G/1 and M^X/G/1 queues with two phases of heterogeneous service under different vacation policies; however, those models assume that each stage of the service is given only to customers already present in the system, whereas our model assumes that one stage of service can be carried out when no customer is present.

The contribution of this paper to the literature is fourfold:

- Formulating a two-dimensional stochastic process that describes the service system outlined above, and obtaining closed-form expressions for the steady-state probabilities and system performance measures.
- Showing that the production and deterioration rates of PSs and the rendering rate of CSs do not affect the stability condition of the service system; this condition is dictated only by the customers' arrival rate and the parameters characterizing an FS.
- Finding that when cost considerations are taken into account, it is not beneficial for the server to utilize all of its idle time to produce PSs, and a limit should be put on the PS inventory, as demonstrated by a concrete example.
- Analyzing the effects of deterioration (spoilage) rate and preservation costs on the decisions of how many PSs to store and how much to invest in preservation technology.

2. Model formulation

We assume that customers arrive to a service system according to a Poisson process with rate λ. An individual customer's service

Please cite this article as: G. Hanukov et al., Performance improvement of a service system via stocking perishable preliminary services, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.10.027

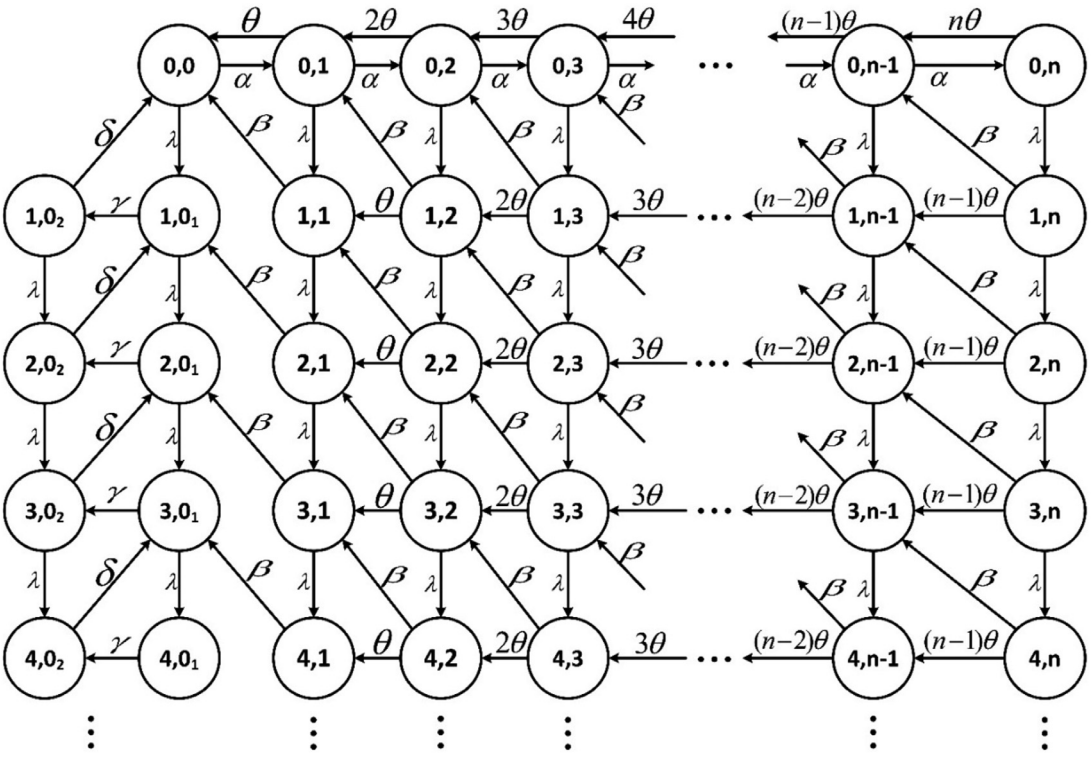


Fig. 1. The system's states and transition-rate diagram.

duration is composed of two independent stages. The first stage (Stage 1) can be completed with or without the customer being present, whereas the second stage (Stage 2) requires the presence of the customer. When the server is idle, it produces an inventory of first-stage services (PSs) for use by future customers. We assume (realistically) that the maximum number of inventoried PSs is limited to n. When the inventory level reaches n and no customer is present, the server stops producing PSs and stays dormant. To facilitate the analysis and focus on our innovative approach, we illustrate our operation method using a single-server queueing system[1] and assume that service times are exponentially distributed. The mean duration of stage 1 is $1/\alpha$ when the customer is absent, and $1/\gamma$ when the customer is present. Note that α does not necessarily equal γ, since the presence of a customer may cause the server to carry out this stage more rapidly in order to satisfy the customer in terms of service duration; alternatively, the customer's presence might interfere with the server's work (e.g., as the customer talks with the server or makes special requests regarding the preparation of the item), thereby lengthening this stage. The mean duration of stage 2 is $1/\beta$ when the server uses a PS, and $1/\delta$ when the two service stages are carried out without interruption. Again, β does not necessarily equal δ, since when using a PS the server might have to carry out setup tasks to retrieve the PS from storage and prepare it for the second stage (e.g., defrosting, unpacking, etc.). We further take into account the process of spoilage of stored PSs and assume that deterioration follows a Markovian process, i.e., the lifetime of a PS is exponentially distributed with parameter θ. Our use of exponential distributions is in line with prior observations (e.g., Cancho, Louzada-Neto, & Barriga, 2011) that the exponential distribution provides a simple, elegant and closed-form solution to many problems in lifetime testing and reliability studies. Indeed, the exponential distribution is commonly used in the literature to describe shelf-life duration of perishable products (see, e.g., Berman & Sapna, 2002, Chao, Iravani, & Savaskan, 2009). When the server is in the middle of producing a PS and a customer arrives, the server stops producing the PS and immediately starts serving the newly-arrived customer.

We formulate the process as a quasi-birth-and-death (QBD) process (see, e.g., Gupta & Selvaraju, 2006) and use matrix geometric methods to analyze the system in steady state (see, e.g., Chakravarthy, 2016; Ma, Liu, & Li, 2013; Van Do, 2015; Zhou, Huang, & Zhang, 2014). Let $L \in \{0,1,2,\ldots\}$ denote the number of customers present in the system, and let $S \in \{0,0_1,0_2,1,2,3\ldots,n\}$ denote the number of PSs in the system, where 0_k, $k \in \{1,2\}$, denotes that there are no PSs in the system, and that the service being provided to the customer at the front of the queue is in its kth stage. Let the system state be (L,S) and define the following steady-state probabilities: $p_{i,j} = \Pr(L = i, S = j)$, where $i = 0$ is followed by $j = 0, 1, 2, 3, \ldots, n$, whereas $i = 1, 2, \ldots$ is followed by $j = 0_1, 0_2, 1, 2, 3, \ldots, n$. The system's states and transition rate diagram are depicted in Fig. 1.

We arrange the system's states in the following order:

$$\{(0,0), (0,1), \ldots, (0,n); (1,0_2), (1,0_1), (1,1), \ldots, (1,n); \ldots; (i,0_2), (i,0_1), (i,1), \ldots, (i,n); \ldots\},$$

and construct its infinitesimal generator matrix (see, e.g., Neuts, 1981, p. 82; Chakravarthy, 2014; Perlman, Elalouf, & Yechiali, 2018),

[1] We have also experimented with a two-server queueing system and obtained closed-form expressions for the system's performance measures for a given n. The analytical expressions are considerably larger than those of a single-server queueing system, and thus are beyond the scope of this paper. Nevertheless, our major results carry over to a two-server system, and we conjecture that they are also valid for multiple servers.

Please cite this article as: G. Hanukov et al., Performance improvement of a service system via stocking perishable preliminary services, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.10.027

denoted by Q, as

$$Q = \begin{pmatrix} B_0 & B_1 & 0 & 0 & 0 & \cdots \\ B_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}.$$

where the matrices B_0, B_1, B_2, A_0, A_1 and A_2 are given in the Appendix.

Denote the vectors of the system's states as $\vec{p}_0 \equiv (p_{0,0}, p_{0,1}, ..., p_{0,n-1}, p_{0,n})$, $\vec{p}_i \equiv (p_{i,0_2}, p_{i,0_1}, p_{i,1}, p_{i,2}, ..., p_{i,n-1}, p_{i,n})$, $i = 1, 2, 3, ...$ Further, let the vector of all probabilities be $\vec{p} \equiv (\vec{p}_0, \vec{p}_1, \vec{p}_2, ...)$, and let $\vec{e}$ be a column vector with all its entries equal to 1. Then, the steady-state probabilities uniquely satisfy

$$\begin{cases} \vec{p}Q = \vec{0} \\ \sum_{i=0}^{\infty} \vec{p}_i \vec{e} = 1 \end{cases} \quad (1)$$

3. Analysis

In this section, we derive the system's stability condition, calculate the system's steady-state probabilities and obtain system performance measures.

3.1. Stability condition

Let

$$A \equiv A_0 + A_1 + A_2 = \begin{pmatrix} -\delta & \delta & 0 & 0 & \cdots & 0 & 0 \\ \gamma & -\gamma & 0 & 0 & \cdots & 0 & 0 \\ 0 & \beta & -\beta & 0 & \cdots & 0 & 0 \\ 0 & 0 & (\beta+\theta) & -(\beta+\theta) & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & \cdots & (\beta+(n-1)\theta) & -(\beta+(n-1)\theta) \end{pmatrix}$$

$\vec{\pi} \equiv (\pi_{0_2}, \pi_{0_1}, \pi_1, \pi_2, ..., \pi_n)$, where $\vec{\pi} A = \vec{0}$ and $\vec{\pi}\,\vec{e} = 1$. Then, the condition for stability is given by (see, e.g., Neuts, 1981):

$$\vec{\pi} A_0 \vec{e} < \vec{\pi} A_2 \vec{e}.$$

Theorem 1. *The system's stability condition is* $\frac{1}{\gamma} + \frac{1}{\delta} < \frac{1}{\lambda}$.

Proof. Straightforward by substituting the expressions of A_0 and A_2.

It follows from Theorem 1 that the stability of the system is not affected by the rates, α and β, of the two stages of a decomposed service, or by θ, the deterioration rate of a PS. This result can be explained as follows: since the number of PSs is bounded, the system shifts gradually to the $S = 0$ level (no inventory of PSs), and then it behaves like an M/G/1 queue with arrival rate λ and mean service time $1/\gamma + 1/\delta$.

3.2. Calculation of the steady-state probabilities

Let R be a matrix of order $(n+2) \times (n+2)$ satisfying

$$A_0 + RA_1 + R^2 A_2 = 0. \quad (2)$$

According to Neuts (1981), the steady-state probabilities satisfy

$$\vec{p}_i = \vec{p}_1 R^{i-1},\ i = 1, 2, 3, ... \quad (3)$$

The explicit entries of the left-hand side of Eq. (2) are given in a Supplementary File. Since there may be several values for each entry in R, only the minimal non-negative value should be taken (Neuts, 1981, p. 82). In most cases, the entries $r_{i,j}$ of the matrix R can be found only by numerical calculations (see Chapter 8 in Latouche & Ramaswami, 1999). A common method for computing the matrix R is by *successive substitutions* (see Harchol-Balter, 2013). However, in our problem, we have successfully obtained closed-form expressions for *all* $r_{i,j}$, $i, j \in \{1, 2, ..., n+2\}$, $n \geq 1$, as given in Theorem 2 below. Such an explicit *complete* solution for the matrix R is rare in the literature and enables large problems to be solved quickly.

Theorem 2. *The entries of the rate matrix* $R \equiv [r_{i,j}]$ *are given by*

$$r_{i,1} = \begin{cases} \lambda/\delta & 1 \leq i \leq 2 \\ \frac{\gamma}{\lambda+\delta} r_{i,2} & 3 \leq i \leq n+2 \end{cases}$$

$$r_{i,2} = \begin{cases} \lambda^2/(\gamma\delta) & i = 1 \\ \lambda(\lambda+\delta)/(\gamma\delta) & i = 2 \\ \frac{(\lambda+\beta+(i-3)\theta)(\lambda+\delta)}{\gamma\delta(\beta+(i-3)\theta)} \left(\delta \sum_{k=3}^{i-1} r_{i,k} r_{k,1} + \beta \sum_{k=3}^{i} r_{i,k} r_{k,3} \right) & 3 \leq i \leq n+2 \end{cases}$$

$$r_{i,j} = \begin{cases} 0 & 3 \leq j \leq n+2,\ i < j \\ \frac{\lambda}{\lambda+\beta+(j-3)\theta} & 3 \leq j = i \leq n+2 \\ \frac{\lambda(\lambda+(j-2)\theta)(\beta+(j-2)\theta)}{(\lambda+\beta+(j-2)\theta)^2(\lambda+\beta+(j-3)\theta)} & 3 \leq j = i-1 \leq n+1 \\ \frac{\begin{pmatrix} \lambda\beta(2\lambda+2\beta+(i+j-5)\theta) \\ +(j-2)\theta(\lambda+\beta+(j-2)\theta)(\lambda+\beta+(i-3)\theta) \end{pmatrix}}{(\lambda+\beta+(j-3)\theta)(\lambda+\beta+(j-2)\theta)(\lambda+\beta+(i-3)\theta)} r_{i,j+1} + \frac{\beta \sum_{k=j+2}^{i-1} r_{i,k} r_{k,j+1}}{\lambda+\beta+(j-3)\theta} & 3 \leq j \leq i-2 \leq n \end{cases}$$

Proof. See Supplementary File.

By Theorem 2, the entries in the main diagonal of R, all the entries above it, and the entries in the first diagonal below it (except $r_{3,2}$) can be calculated directly by the model parameters. Other entries are calculated recursively diagonally starting from $r_{n+2,n}$ to $r_{5,3}$, then starting from $r_{n+2,n-1}$ to $r_{6,3}$, and so on until $r_{n+2,3}$. Then, we calculate the entries from $r_{3,2}$ down to $r_{n+2,2}$, and finally we calculate the entries from $r_{3,1}$ down to $r_{n+2,1}$. For example, when $n = 2$, the R matrix is:

Please cite this article as: G. Hanukov et al., Performance improvement of a service system via stocking perishable preliminary services, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.10.027

$$R=\begin{pmatrix} \frac{\lambda}{\delta} & \frac{\lambda^2}{\gamma\delta} & 0 & 0 \\ \frac{\lambda}{\delta} & \frac{\lambda(\lambda+\delta)}{\gamma\delta} & 0 & 0 \\ \frac{\lambda^2}{\delta(\lambda+\beta)} & \frac{\lambda^2(\lambda+\delta)}{\gamma\delta(\lambda+\beta)} & \frac{\lambda}{\lambda+\beta} & 0 \\ \frac{\lambda^2(\lambda+\theta)(2\beta+\lambda+\theta)}{\delta(\lambda+\beta)(\lambda+\beta+\theta)^2} & \frac{\lambda^2(\lambda+\theta)(2\beta+\lambda+\theta)(\lambda+\delta)}{\gamma\delta(\lambda+\beta)(\lambda+\beta+\theta)^2} & \frac{\lambda(\lambda+\theta)(\beta+\theta)}{(\lambda+\beta)(\lambda+\beta+\theta)^2} & \frac{\lambda}{\lambda+\beta+\theta} \end{pmatrix}$$

In order to calculate, via Eq. (3), $p_{0,0}$, $p_{i,0_k}$, $i=1,2,3,\ldots$, $k=1,2$ and $p_{i,j}$, $i=0,1,2,\ldots$, $j=1,2,\ldots,n$, we first have to obtain the vector of boundary probabilities $\vec{p}_0$ and the vector $\vec{p}_1$ by solving (part of the set in Eq. (1)):

$$\begin{aligned} &\vec{p}_0B_0+\vec{p}_1B_2=\vec{0} \\ &\vec{p}_0B_1+\vec{p}_1[A_1+RA_2]=\vec{0} \\ &\vec{p}_0\vec{e}+\vec{p}_1[I-R]^{-1}\vec{e}=1. \end{aligned} \tag{4}$$

4. Performance measures

Since n (the maximal capacity of PSs) is the only parameter that dictates the complexity of the queueing-inventory system, we define the following performance measures as functions of n. For a given value of n, let $L(n)$ and $L_q(n)$ denote, respectively, the mean number of customers in the system and the mean number of customers in queue. Similarly, let $W(n)$ and $W_q(n)$ denote, respectively, a customer's mean sojourn time in the system and a customer's mean sojourn time in queue; let $S(n)$ and $S_q(n)$ denote, respectively, the mean number of PSs in the system and the mean number of PSs in inventory; and let $T(n)$ and $T_q(n)$ denote, respectively, the mean duration of time that a PS resides in the system and the mean duration of time that it resides in inventory.

Using Eq. (3), we have:

$$L(n)=\sum_{i=1}^{\infty} i(\vec{p}_i\vec{e})=\sum_{i=1}^{\infty} i(\vec{p}_1R^{i-1}\vec{e})=\vec{p}_1\left(\sum_{i=1}^{\infty} iR^{i-1}\right)\vec{e}.$$

Since $\sum_{i=1}^{\infty} iR^{i-1}=([I-R]^{-1})^2=[I-R]^{-2}$, then

$$L(n)=\vec{p}_1[I-R]^{-2}\vec{e}. \tag{5}$$

Since $\vec{p}_0\vec{e}$ is the probability that the system is empty of customers, we readily obtain

$$L_q(n)=L(n)-(1-\vec{p}_0\vec{e}). \tag{6}$$

By Little's law,

$$W(n)=L(n)/\lambda, \tag{7}$$

$$W_q(n)=L_q(n)/\lambda. \tag{8}$$

In order to obtain $S(n)$, we define two column vectors $\vec{v}\equiv(0,\,1,\,2,\ldots,n)^T$ and $\vec{w}\equiv(0,\,0,\,1,\,2,\ldots,n)^T$. Thus,

$$\begin{aligned} S(n)&=\sum_{i=0}^{\infty}\sum_{j=1}^{n} jp_{i,j}=\vec{p}_0\vec{v}+\sum_{i=1}^{\infty}\vec{p}_i\vec{w} \\ &=\vec{p}_0\vec{v}+\vec{p}_1\left(\sum_{i=1}^{\infty}R^{i-1}\right)\vec{w}=\vec{p}_0\vec{v}+\vec{p}_1[I-R]^{-1}\vec{w}. \end{aligned} \tag{9}$$

Similarly, in order to get $S_q(n)$, we define a column vector $\vec{u}\equiv(0,\,0,\,0,\,1,\,2,\ldots,n-1)^T$ for $n\geq 1$, leading to

$$\begin{aligned} S_q(n)&=\sum_{j=1}^{n} jp_{0,j}+\sum_{i=1}^{\infty}\sum_{j=1}^{n}(j-1)p_{i,j} \\ &=\vec{p}_0\vec{v}+\sum_{i=1}^{\infty}\vec{p}_i\vec{u}=\vec{p}_0\vec{v}+\vec{p}_1\left(\sum_{i=1}^{\infty}R^{i-1}\right)\vec{u}=\vec{p}_0\vec{v}+\vec{p}_1[I-R]^{-1}\vec{u}. \end{aligned} \tag{10}$$

Although the server's PS production rate is α, the server generates PSs only when there are no customers and the inventory level is less than n. Therefore, the average production rate of PSs is

$$\bar{\alpha}(n)=\alpha(\vec{p}_0\vec{e}-p_{0,n}). \tag{11}$$

Using Little's law, we obtain:

$$T(n)=S(n)/\bar{\alpha}(n), \tag{12}$$

$$T_q(n)=S_q(n)/\bar{\alpha}(n). \tag{13}$$

The average deterioration rate of PSs is $\bar{\theta}(n)=\sum_{j=1}^{n} j\theta p_{0,j}+\sum_{i=1}^{\infty}\sum_{j=1}^{n}(j-1)\theta p_{i,j}$. By Eq. (10),

$$\bar{\theta}(n)=\theta S_q(n). \tag{14}$$

Indeed, since each PS deteriorates independently of other PSs, the mean rate of deterioration is proportional to the mean inventory level of PSs.

It is interesting to calculate the proportion of customers who wait only for the second service-stage, since the type of actual service (second stage only or two stages) may affect customers' satisfaction. We claim:

Proposition 1. *The proportion of customers who wait only for the second service-stage is* $\eta(n)=\frac{\bar{\alpha}(n)-\bar{\theta}(n)}{\lambda}$.

Proof. The claim readily follows from the fact that the difference $\bar{\alpha}(n)-\bar{\theta}(n)$ is the rate of PSs being used. Since the capacity of PSs is finite and equals n, the following relation holds: $\bar{\alpha}(n)-\bar{\theta}(n)<\lambda$; otherwise, the inventory of PSs will diverge to infinity in the long run, which is a contradiction. □

The traditional M/M/1 queue can be captured by our model as a degenerate case in which $\beta\to\infty$, $\delta\to\infty$ (i.e., the second stage of service (CS) has no duration, implying that a full service (FS) is given in the first stage when a customer is present) and $\alpha=0$ (i.e., when the server is idle he has no ability to prepare PSs). A detailed proof of this claim is given in Appendix 2.

5. Economic analysis

In addition to the operational performance measures defined in the previous section, we consider an integrative economic measure to evaluate the performance of the system, namely, the system's long-run cost rate. This measure is commonly used in practice and in literature to evaluate the efficiency of service

Table 1
The values of $G(n|\theta)$ for $n \in \{0,1,2,\ldots,20\}$ with $\theta \in \{0,0.05,0.10,\ldots,0.50\}$.

$n\backslash\theta$	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0	9.867	9.867	9.867	9.867	9.867	9.867	9.867	9.867	9.867	9.867	9.867
1	9.277	8.964	8.817	8.737	8.690	8.662	8.646	8.638	8.635	8.636	8.640
2	9.015	8.407	8.132	7.989	7.913	7.873	7.857	7.856	7.865	7.882	7.904
3	**8.997**	8.107	7.715	7.521	7.425	7.384	7.377	7.391	7.420	7.459	7.506
4	9.172	**8.008**	7.507	7.268	7.158	7.120	7.126	7.159	7.211	7.275	**7.348**
5	9.502	8.071	**7.464**	**7.183**	**7.062**	**7.029**	**7.048**	**7.101**	**7.175**	**7.263**	7.360
6	9.960	8.262	7.552	7.231	7.098	7.069	7.101	7.171	7.265	7.374	7.492
7	10.523	8.559	7.746	7.383	7.238	7.211	7.253	7.337	7.446	7.571	7.705
8	11.170	8.942	8.024	7.618	7.458	7.430	7.479	7.572	7.692	7.828	7.973
9	11.889	9.393	8.370	7.919	7.741	7.708	7.759	7.857	7.983	8.124	8.273
10	12.665	9.902	8.770	8.271	8.072	8.031	8.078	8.176	8.301	8.442	8.589
11	13.489	10.456	9.214	8.664	8.440	8.386	8.425	8.517	8.637	8.770	8.909
12	14.353	11.048	9.694	9.090	8.837	8.765	8.791	8.871	8.979	9.100	9.226
13	15.249	11.672	10.203	9.541	9.255	9.160	9.168	9.231	9.321	9.424	9.532
14	16.172	12.321	10.735	10.013	9.689	9.567	9.551	9.590	9.658	9.739	9.824
15	17.117	12.991	11.286	10.500	10.134	9.979	9.934	9.945	9.986	10.040	10.100
16	18.081	13.678	11.851	10.998	10.587	10.394	10.314	10.293	10.302	10.327	10.359
17	19.059	14.379	12.429	11.506	11.044	10.808	10.689	10.631	10.605	10.598	10.601
18	20.050	15.092	13.017	12.019	11.502	11.219	11.056	10.957	10.894	10.854	10.829
19	21.052	15.815	13.613	12.537	11.961	11.625	11.414	11.271	11.170	11.096	11.043
20	22.061	16.546	14.214	13.057	12.417	12.024	11.761	11.573	11.433	11.326	11.246

systems (e.g., Huang, Carmeli, & Mandelbaum, 2015; Wang, Zhang, & Huang, 2017; Xu et al. 2015; Yang and Wu 2015). Several cost components that should be considered in this integrative measure with respect to our model are: (i) sojourn cost of customers in the system; (ii) holding cost of PSs in inventory; (iii) spoilage cost of PSs; and (iv) a capacity-technology cost associated with maintaining a certain level of PS deterioration rate for a given PS capacity (reflecting the idea that different technologies are associated with different deterioration rates). In order to obtain insights into the structure of the optimal solution, we examine two optimization models where both models are comprised of the above four components.

In the first (basic) model, the server controls only the PS capacity n, where the components are calculated as follows: the first is proportional (with coefficient c) to $L(n)$, the mean number of customers in the system; the second is proportional (with coefficient h) to $S_q(n)$, the mean number of inventoried PSs; the third is proportional (with coefficient d) to $\bar{\theta}(n)$, the average deterioration rate of PSs; and the fourth is simultaneously proportional to the PS capacity n (with coefficient κ_1) and is a hyperbolic (convex) decreasing function (in line with the law of diminishing returns) of a positively-shifted deterioration rate $\theta+\kappa_2$ (where κ_2 is a constant with the same measure-unit as θ that ensures a finite value for $\kappa_1 n/(\theta+\kappa_2)$ at $\theta=0$). The fourth cost component, $\kappa_1 n/(\theta+\kappa_2)$, reflects, for example, energy requirements in a cooling system, which become costlier as the size of the system increases. Consequently, the expected cost function for a given θ is

$$\begin{aligned} G(n|\theta) &\equiv cL(n)+hS_q(n)+d\bar{\theta}+\kappa_1 n/(\theta+\kappa_2) \\ &= cL(n)+(h+\theta d)S_q(n)+\kappa_1 n/(\theta+\kappa_2). \end{aligned} \tag{15}$$

In the second (extended) model, both the PS capacity n and the PS deterioration rate θ can be controlled simultaneously by the decision maker. This means that the decision maker may choose among different PS preservation technologies (e.g., cooling systems), where a lower value of θ is associated with a more advanced technology (e.g., more temperature sensors with better accuracy), which results in a higher operational cost. To emphasize that θ is a decision variable in this model (in addition to n), we include it as an argument of all performance measures depending on it. Thus, the total average cost rate to be minimized is

$$G(n,\theta)=cL(n,\theta)+(h+\theta d)S_q(n,\theta)+\kappa_1 n/(\theta+\kappa_2). \tag{16}$$

In order to demonstrate the applicability and usefulness of our models, we provide a practical example from the fast food industry. Consider a single server in a coffee shop that sells breakfasts comprising a sandwich and different types of coffee. Customers arrive according to a Poisson process with rate $\lambda=8$ per hour. Stage 1 of the service is preparing the sandwich, which takes on average 4 minutes when the customer is either present or absent ($\alpha=\gamma=15$), and includes slicing bread, adding chopped vegetables, cheese, etc., and wrapping it up. Stage 2 includes preparing the coffee according to the customer's requirements (e.g., macchiato, espresso, cappuccino, etc.), which takes on average 2 minutes ($\beta=\delta=30$). In addition, the sojourn cost of a customer is estimated at $c=\$3$ per hour. The cost of holding a prepared sandwich in cool storage is estimated at $h=\$0.05$ per hour, and the cost incurred by disposing of a sandwich that is not suitable for selling is estimated at $d=\$1.5$. Finally, we use $k_1=\$0.1$ per square hour (so that the units of $\kappa_1 n/(\theta+\kappa_2)$ are \$ per hour) and $\kappa_2=1$ unit per hour (the same type of unit used for θ).

5.1. The basic model

Using Eq. (15), we wish to investigate the effect of n on $G(n|\theta)$ and find its optimal value for various values of θ. Table 1 presents the values of $G(n|\theta)$ for all combinations of $n=\{0,1,2,\ldots,20\}$ with $\theta=\{0,0.05,0.10,\ldots,0.50\}$, where the optimal values of n for each value of θ are printed in bold, and Fig. 2 graphically illustrates $G(n|\theta)$.

Analysis of $G(n+1|\theta)-G(n|\theta)$ for the above sets of n and θ shows that $G(n|\theta)$ is quasi-convex over the PS capacity n. Fig. 3 depicts (i) the optimal PS capacity, and (ii) the cost reduction in percentage obtained by adopting the proposed operational method of producing and storing an optimal capacity of PSs, as compared with a system in which no PS is stored at all, for various values of the deterioration rate. We find that a higher deterioration rate of PSs first results in a higher optimal PS capacity, followed by a capacity stabilization, and then in a reduction of its value

Please cite this article as: G. Hanukov et al., Performance improvement of a service system via stocking perishable preliminary services, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.10.027

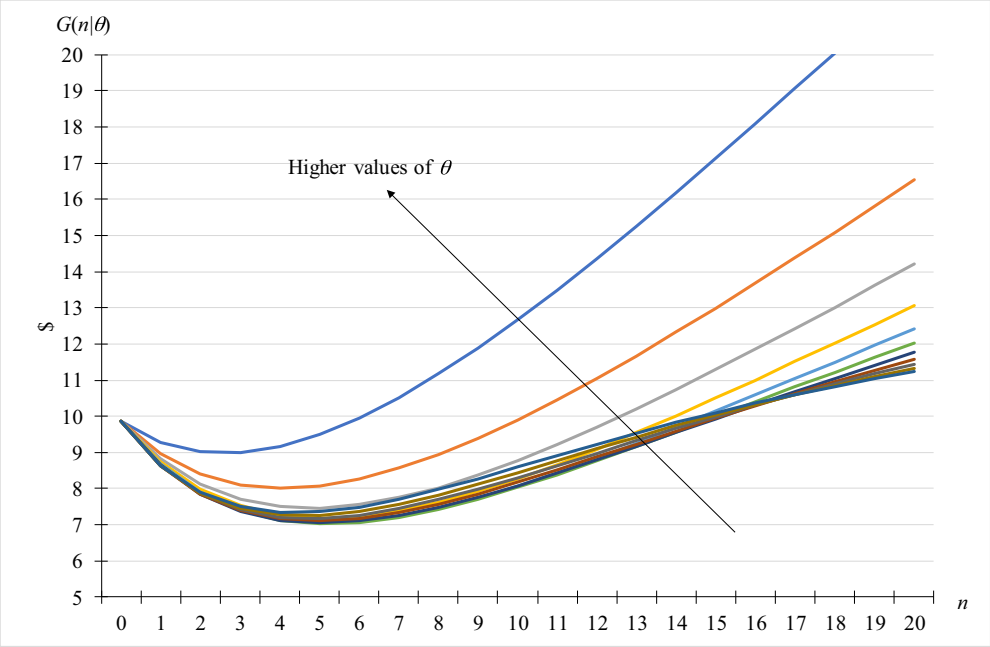


Fig. 2. $G(n|\theta)$ for $n \in \{0,1,2,\ldots,20\}$ with $\theta \in \{0,0.05,0.10,\ldots,0.50\}$.

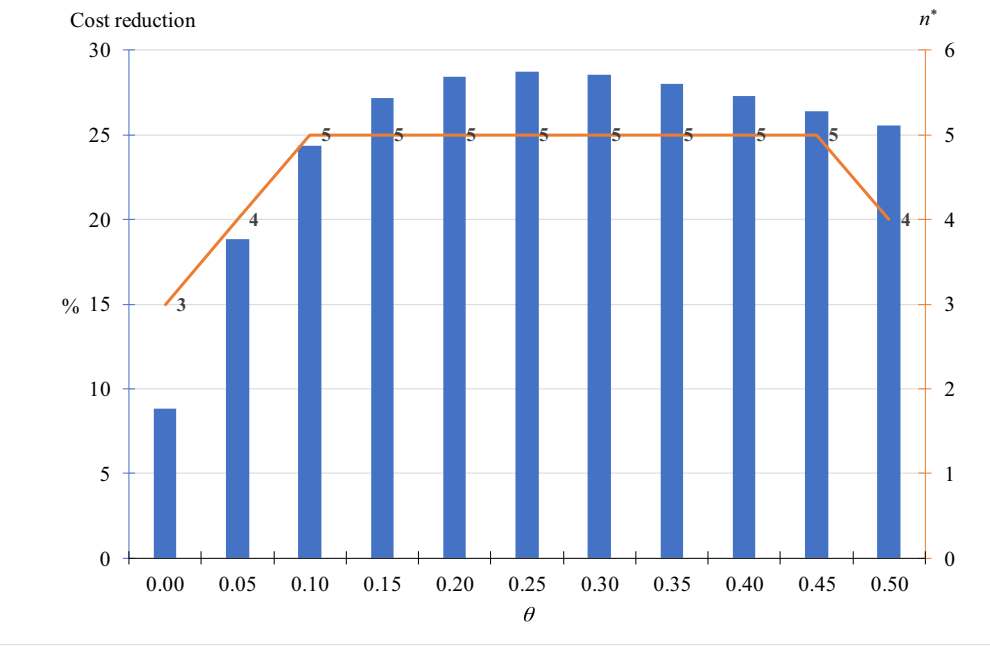


Fig. 3. Optimal capacity n^* and cost reduction in percentage of $G(n^*|\theta)$ compared to $G(0|\theta)$ for various values of θ.

(as depicted by the solid line). A similar phenomenon is observed with regard to the cost reduction when comparing the system to one that does not include the capacity to inventory PSs (as depicted by the bar diagram).

5.2. The extended model

Using Eq. (16), we wish to investigate the effects of both n and θ on $G(n,\theta)$. Table 1 also presents the values of $G(n,\theta)$ for all combinations of $n = \{0, 1, 2, \ldots, 20\}$ with $\theta = \{0, 0.05, 0.10, \ldots, 0.50\}$. The underlined value (7.029) indicates the minimal overall cost rate obtained for $n = 5$ and $\theta = 0.25$. For each θ, denote by $G^*(\theta)$ the minimal cost along the corresponding column in Table 1, and let $G^* \equiv \min_\theta\{G^*(\theta)\}$. Then, Table 2 gives the percentage reduction in the total average cost rate obtained by controlling θ, calculated as $\xi \equiv (G^*(\theta) - G^*)/G^*(\theta) \times 100$. For example, when $\theta = 0$ the percentage improvement is $(8.997 - 7.029)/8.997 = 21.87\%$.

Fig. 4 graphically illustrates $G(n,\theta)$ as a surface chart. Analyzing the numerical results of table 1, we observe that: (i) $G(n,\theta)$ is a quasi-convex function with a minimum; (ii) for low values of θ (i.e., $\theta < 0.1$), a higher value of θ results in a higher value of n^*; (iii) for medium values of θ (i.e., $0.1 \le \theta < 0.45$), a higher value of θ does not affect the value of n^*; and (iv) for high values of θ (i.e., $0.45 \le \theta < 0.\ 5$), a higher value of θ results in a lower value of n^*.

Please cite this article as: G. Hanukov et al., Performance improvement of a service system via stocking perishable preliminary services, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.10.027

Table 2
Values of the percentage reduction in the total average cost rate ξ for $\theta \in \{0,0.05,0.10,\ldots,0.50\}$.

θ	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ξ	21.87	12.23	5.83	2.14	0.47	0.00	0.27	1.01	2.03	3.22	4.34

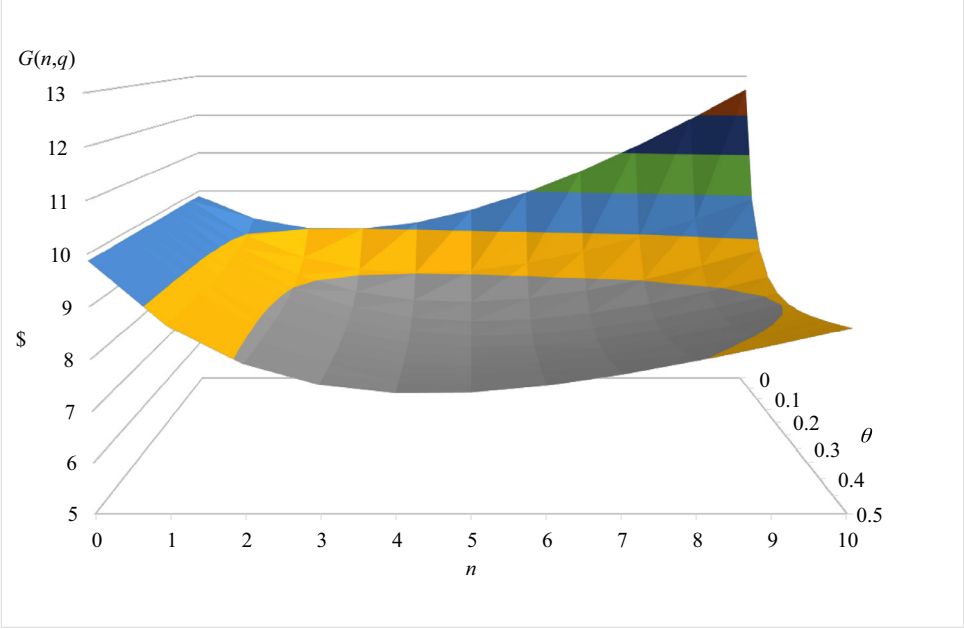


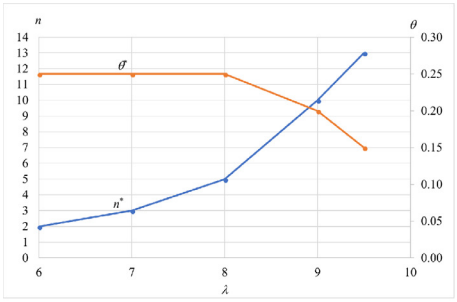
Fig. 4. $G(n,\theta)$ for $n \in \{0,1,2,\ldots,10\}$ with $\theta \in \{0,0.05,0.10,\ldots,0.50\}$.

5.3. Sensitivity analysis for the extended model

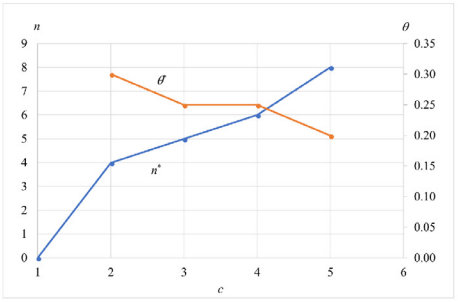
In order to investigate the effect of each parameter on the optimal decisions n^* and θ^*, we use the numerical example in 5.2 as a base case and calculate (n^*,θ^*) for various values of the parameters, where each time we change only the value of a single parameter (selecting values both below and above the base-case value), keeping the other parameter values constant. The following parameter values are used: $\lambda \in \{6, 7, 8, 9, 9.5\}$, $\alpha \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, $\beta \in \{20, 25, 30, 35, 40\}$, $\gamma \in \{13, 14, 15, 16, 17\}$, $\delta \in \{20, 25, 30, 35, 40\}$, $c \in \{1, 2, 3, 4, 5\}$, $d \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$, $h \in \{0.03, 0.04, 0.05, 0.06, 0.07\}$, $\kappa_1 \in \{0.02, 0.06, 0.10, 0.14, 0.18\}$ and $\kappa_2 \in \{0.02, 0.06, 0.10, 0.14, 0.18\}$. We note that in the case of parameter α, we use ten values instead of five, because changes to the value of this parameter produce non-monotonic effects, which are revealed for α values that are considerably larger than that of the base case. The results are presented in Fig. 5(a)–(j).

Fig. 5(a) shows that a higher arrival rate requires both a higher PS capacity and a lower PS deterioration rate, which work simultaneously to enable the server to fulfill the higher demand. As expected, Fig. 5(b) shows that a higher sojourn cost rate of a customer requires a higher PS capacity and a lower PS deterioration rate, since in this case it becomes a higher priority to reduce the customer queue by using PSs than to ensure that PSs can be stored for long durations. Fig. 5(c) points to an interesting result: (i) Within a range of low values of α, an increase in the value of this parameter requires higher PS capacity and a lower PS deterioration rate. These trends can be explained by the server's efficiency in producing PSs: when the server is more efficient, a larger inventory can be produced during the time in which no customers are in the system, and thus, in order to prevent a large number of spoiled units, a better preservation technology should be used. (ii) Within a range of medium values of α, it is not beneficial to further decrease the deterioration rate of PSs due to the law of diminishing returns, and thus there is no justification to increase further the PS capacity when α increases. (iii) For high values of α, the server prepares PSs so rapidly that some reduction in the PS capacity is beneficial, as it prevents the accumulation of a large average inventory that results in high holding costs. In this case, in line with the explanation of (ii), it is not beneficial to further decrease the deterioration rate of PSs. Fig. 5(d) shows that a higher CS rendering rate requires a higher PS capacity and a lower PS deterioration rate; however, the latter values stabilize beyond a certain value of the CS rendering rate, because for higher values of β the server will quickly serve the customers and will again have idle time to produce fresh PSs. Fig. 5(e) shows that a higher value of γ (service rate of first-stage service in the presence of the customer) requires a lower PS capacity and allows a higher PS deterioration rate, since it becomes less beneficial to produce and store PSs compared with carrying out the full service in the presence of the customer. Fig. 5(f) shows the same trends as Fig. 5(e) with respect to the service rate of the second stage, for the same reasons. Fig. 5(g) shows that a higher spoilage cost per unit results in both a lower PS capacity and a lower PS deterioration rate, effectively reflecting the need to employ as many means as possible to reduce the risk of spoilage. Fig. 5(h) shows that a higher holding cost rate per unit does not affect the required PS capacity and PS deterioration rate; this observation can be explained by the low contribution of this cost component to the total cost in our numerical example. Fig. 5(i) shows that a higher scale coefficient of the spoilage cost component requires a lower PS capacity until the latter value stabilizes, and it allows a higher PS deterioration rate (the relation between κ_1 and θ^* is approximately linear). Fig. 5(j) shows that a higher value of κ_2 either does not change the optimal PS capacity or results in a slightly higher value, whereas it requires a lower PS deterioration rate (again, the relation between κ_2 and θ^* is approximately linear).

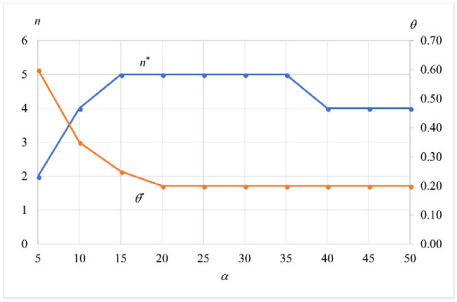
Please cite this article as: G. Hanukov et al., Performance improvement of a service system via stocking perishable preliminary services, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.10.027



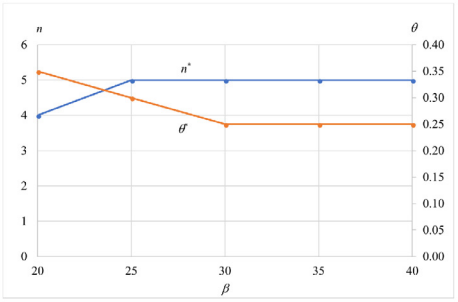
(a) α=γ=15, β=δ=30, c=3, d=1.5, h=0.05, κ_1=κ_2=0.1



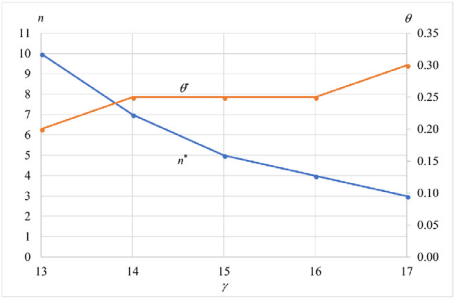
(b) λ=8, α=γ=15, β=δ=30, d=1.5, h=0.05, κ_1=κ_2=0.1



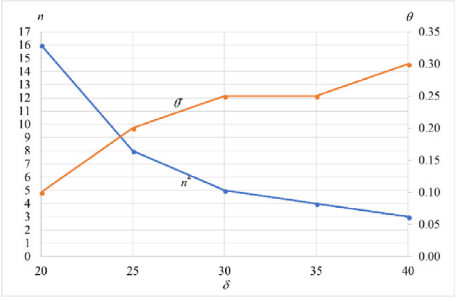
(c) λ=8, γ=15, β=δ=30, c=3, d=1.5, h=0.05, κ_1=κ_2=0.1



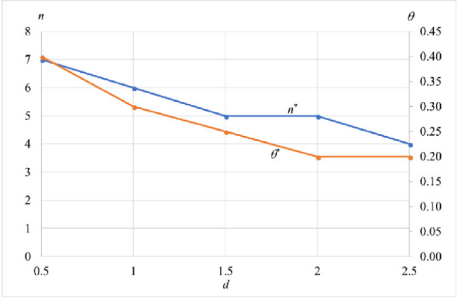
(d) λ=8, α=γ=15, δ=30, c=3, d=1.5, h=0.05, κ_1=κ_2=0.1



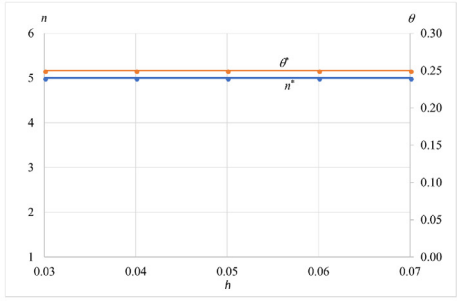
(e) λ=8, α=15, β=δ=30, c=3, d=1.5, h=0.05, κ_1=κ_2=0.1



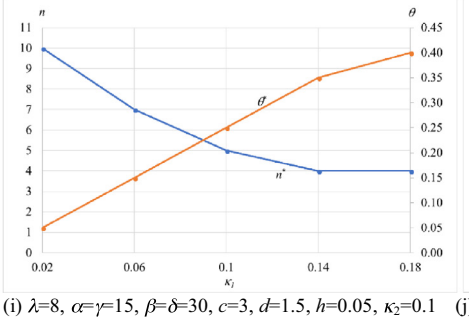
(f) λ=8, α=γ=15, β=30, c=3, d=1.5, h=0.05, κ_1=κ_2=0.1



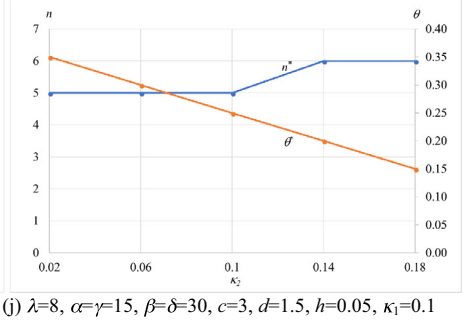
(g) λ=8, α=γ=15, β=δ=30, c=3, h=0.05, κ_1=κ_2=0.1



(h) λ=8, α=γ=15, β=δ=30, c=3, d=1.5, κ_1=κ_2=0.1



(i) λ=8, α=γ=15, β=δ=30, c=3, d=1.5, h=0.05, κ_2=0.1



(j) λ=8, α=γ=15, β=δ=30, c=3, d=1.5, h=0.05, κ_1=0.1

Fig. 5. Sensitivity analysis: n^* and θ^* for various parameter values.

6. Conclusions

In some service systems, servers' idle time can be utilized to produce and store preliminary services (PSs) for future incoming customers. We have modeled and analyzed such a system, introducing an important innovation: the consideration that PSs might spoil while being held in storage, which is a crucial factor in ensuring applicability of the model to the fast food industry. Our system comprises a single-server model, where customers' arrival, the duration of each service stage, and the deterioration rate of stored PSs each follow a Markovian process. Assigning costs to various performance measures of the system, we have sought to identify optimal values of decision variables, including the maximum capacity of inventoried PSs, and the optimal PS deterioration rate (reflecting the decision maker's investment in preservation technologies).

For our analysis, we constructed a two-dimensional state space that considers the number of customers in the system, the number of stored PSs and the stage of the service. Using matrix geometric methods, we obtained the steady-state probabilities and several performance measures analytically, an achievement that enables large systems to be analyzed in a reasonable time. We showed that the PS production and deterioration rates and the CS rendering rate do not affect the stability condition of the service system, which is dictated only by the customers' arrival rate and the parameters characterizing an FS. Specifically, the stability condition is identical to that of an M/G/1 queue, which means that the maximal arrival rate of customers that the server can handle does not differ between the proposed model and a system without PSs. We used the proposed model to carry out an economic analysis of the two-stage service system described, and to compare its performance with that of a system without PSs ($n=0$). Two cost models were evaluated: basic and extended, which differed in the ability of the decision maker to control the deterioration rate and its associated cost.

Numerical analysis reveals that the cost function is apparently quasi-convex over the PS capacity, indicating that it is not beneficial for the server to utilize all of its idle time to produce PSs; an efficient line-search can be used to obtain the optimum. In addition, we find that the optimal PS capacity is not monotonic in the deterioration rate, but first increases, then stabilizes and finally decreases. We analyzed the sensitivity of the optimal values of each of the decision variables to each of the model parameters, and found that for some parameters they are monotonic, whereas for others they stabilize or even change direction.

We suggest several possible directions for further research in this domain. One option is to extend the above two-stage operational model to a multi-server system or to a system with a limited customer queue capacity. Another interesting direction would be to investigate a system with rates/durations that follow general distributions instead of Poisson/exponential distributions. Analysis of such a framework would require the use of simulations due to lack of mathematical tractability. The Markovian model proposed in this work could be evaluated as an approximation for the general case. Finally, modeling a system with different customer attitudes towards using preliminary services is a potentially fruitful avenue for future investigation in the domain of behavioral operations research.

Acknowledgment

This research was supported by the Israel Science Foundation (grant No. 1448/17).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2018.10.027.

Appendix 1

$$B_0=\begin{pmatrix} -(\alpha+\lambda) & \alpha & 0 & 0 & \cdots & 0 & 0 \\ \theta & -(\alpha+\theta+\lambda) & \alpha & 0 & & 0 & 0 \\ 0 & 2\theta & -(\alpha+2\theta+\lambda) & \alpha & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & & & -(\alpha+(n-1)\theta+\lambda) & \alpha \\ 0 & 0 & 0 & 0 & \cdots & n\theta & -(\lambda+n\theta) \end{pmatrix}$$

$$B_1=\begin{pmatrix} 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & & \lambda & 0 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix},\quad B_2=\begin{pmatrix} \delta & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \beta & 0 & & 0 & 0 \\ 0 & \beta & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta & 0 \end{pmatrix},\quad A_0=\begin{pmatrix} \lambda & 0 & \cdots & 0 & 0 \\ 0 & \lambda & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & \lambda & 0 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix},$$

$$A_1=\begin{pmatrix} -(\delta+\lambda) & 0 & 0 & 0 & \cdots & 0 & 0 \\ \gamma & -(\gamma+\lambda) & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -(\beta+\lambda) & 0 & \cdots & 0 & 0 \\ 0 & 0 & \theta & -(\beta+\theta+\lambda) & & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & \cdots & (n-1)\theta & -(\beta+(n-1)\theta+\lambda) \end{pmatrix},\text{ and}$$

$$A_2 = \begin{pmatrix} 0 & \delta & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \beta & 0 & \cdots & 0 & 0 \\ 0 & 0 & \beta & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \beta & 0 \end{pmatrix}.$$

Appendix 2

By substituting $\delta \to \infty$ in Theorem 1, we get the stability condition of the traditional M/M/1 queue with mean arrival rate λ and mean service time $1/\gamma$, i.e., $\lambda < \gamma$. As for the boundary probabilities, by substituting $\alpha = 0$ and $\delta \to \infty$ in Eq. (4) and solving the set of equations, we get $\vec{p}_0 = (1 - \lambda/\gamma, 0, 0, ..., 0)$ and $\vec{p}_1 = (0, (1 - \frac{\lambda}{\gamma})\frac{\lambda}{\gamma}, 0, 0, ..., 0)$, which represent the probabilities of no customer and one customer in the system, respectively, in the above M/M/1 queue. By substituting $\beta \to \infty$, $\delta \to \infty$ in Theorem, 2 we get $r_{i,j} = \begin{cases} \lambda/\gamma & i = j = 2 \\ 0 & \text{otherwise} \end{cases}$. This implies that the element $[i,j]$ in matrix $[I - R]^{-1}$ is:

$$[I-R]^{-1}{}_{[i,j]} = \begin{cases} 1 & i = j = 1, 3, 4, ..., n+2 \\ \gamma/(\gamma - \lambda) & i = j = 2 \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

and element $[i,j]$ in matrix $[I - R]^{-2}$ is:

$$[I-R]^{-2}{}_{[i,j]} = \begin{cases} 1 & i = j = 1, 3, 4, ..., n+2 \\ \gamma^2/(\gamma - \lambda)^2 & i = j = 2 \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

Substituting Eq. (2.2) in Eqs. (5) and (6) yields $L(n) = \lambda/(\gamma - \lambda)$, $L_q(n) = \lambda^2/(\gamma(\gamma - \lambda))$; these expressions correspond to the number of customers in the system and in queue, respectively, in the above M/M/1 queue. Similarly, by substituting Eq. (2.1) in Eqs. (9) and (10), we get $S(n) = 0$ and $S_q(n) = 0$, which means no PSs at all.

References

Adacher, L., & Cassandras, C. G. (2014). Lot size optimization in manufacturing systems: The surrogate method. *International Journal of Production Economics, 155*, 418–426.
Altendorfer, K., & Minner, S. (2015). Influence of order acceptance policies on optimal capacity investment with stochastic customer required lead times. *European Journal of Operational Research, 243*(2), 555–565.
Avşar, Z. M., & Zijm, W. H. (2014). Approximate queueing models for capacitated multi-stage inventory systems under base-stock control. *European Journal of Operational Research, 236*(1), 135–146.
Avinadav, T., & Arponen, T. (2009). An EOQ model for items with a fixed shelf-life and a declining demand rate based on time-to-expiry. *Asia-Pacific Journal of Operational Research, 26*(6), 759–767.
Avinadav, T., Herbon, A., & Spiegel, U. (2013). Optimal inventory policy for a perishable item with demand function sensitive to price and time. *International Journal of Production Economics, 144*, 497–506.
Avinadav, T., Herbon A. , A., & Spiegel, U. (2014). Optimal ordering and pricing policy for demand functions that are separable into price and inventory age. *International Journal of Production Economics, 155*, 406–417.
Avinadav, T., Chernonog, T., Lahav, Y., & Spiegel, U. (2017). Dynamic pricing and promotion expenditures in an EOQ model of perishable items. *Annals of Operations Research, 248*, 75–91.
Berk, E., & Gürler, Ü. (2008). Analysis of the (Q, r) inventory model for perishables with positive lead times and lost sales. *Operations Research, 56*(5), 1238–1246.
Berman, O., & Sapna, K. P. (2002). Optimal service rates of a service facility with perishable inventory items. *Naval Research Logistics, 49*(5), 464–482.
Boxma, O. J., Schlegel, S., & Yechiali, U. (2002). A note on the M/G/1 queue with a waiting server, timer and vacations. *American Mathematical Society Translations, Series, 2*(207), 25–35.
Cancho, V. G., Louzada-Neto, F., & Barriga, G. D. (2011). The Poisson-exponential lifetime distribution. *Computational Statistics & Data Analysis, 55*(1), 677–686.
Chakravarthy, S. R. (2014). A multi-server queueing model with server consultations. *European Journal of Operational Research, 233*(3), 625–639.
Chakravarthy, S. R. (2016). Queueing models with optional cooperative services. *European Journal of Operational Research, 248*(3), 997–1008.
Chao, G. H., Iravani, S. M., & Savaskan, R. C. (2009). Quality improvement incentives and product recall cost sharing contracts. *Management Science, 55*(7), 1122–1138.
Chao, X., Gong, X., Shi, C., & Zhang, H. (2015). Approximation algorithms for perishable inventory systems. *Operations Research, 63*(3), 585–601.
Chebolu-Subramanian, V., & Gaukler, G. M. (2015). Product contamination in a multi-stage food supply chain. *European Journal of Operational Research, 244*(1), 164–175.
Chen, X., Pang, Z., & Pan, L. (2014). Coordinating inventory control and pricing strategies for perishable products. *Operations Research, 62*(2), 284–300.
Chernonog, T., & Avinadav, T. (2017). *Pricing and advertising in a supply chain of perishable products under asymmetric information*. Bar-Ilan University Working paper.
Choudhury, G. (2007). A two phase batch arrival retrial queueing system with Bernoulli vacation schedule. *Applied Mathematics and Computation, 188*, 1455–1466.
Choudhury, G. (2008). Steady state analysis of an M/G/1 queue with linear retrial policy and two phase service under Bernoulli vacation schedule. *Applied Mathematical Modelling, 32*, 2480–2489.
Choudhury, G., & Deka, M. (2012). A single server queueing system with two phases of service subject to server breakdown and Bernoulli vacation. *Applied Mathematical Modelling, 36*, 6050–6060.
Choudhury, G., & Madan, K. (2005). A two-stage batch arrival queueing system with a modified Bernoulli schedule vacation under N-policy. *Mathematical and Computer Modelling, 42*, 71–85.
Choudhury, G., Tadj, L., & Paul, M. (2007). Steady state analysis of an M x/G/1 queue with two phase service and Bernoulli vacation schedule under multiple vacation policy. *Applied Mathematical Modelling, 31*, 1079–1091.
Cooper, W. L. (2001). Pathwise properties and performance bounds for a perishable inventory system. *Operations Research, 49*(3), 455–466.
Dye, C. Y. (2013). The effect of preservation technology investment on a non-instantaneous deteriorating inventory model. *Omega, 41*(5), 872–880.
Dye, C. Y., & Hsieh, T. P. (2012). An optimal replenishment policy for deteriorating items with effective investment in preservation technology. *European Journal of Operational Research, 218*(1), 106–112.
Guha, D., Goswami, V., & Banik, A. D. (2016). Algorithmic computation of steady-state probabilities in an almost observable GI/M/c queue with or without vacations under state dependent balking and reneging. *Applied Mathematical Modelling, 40*(5), 4199–4219.
Guo, P., & Zhang, Z. G. (2013). Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. *Manufacturing & Service Operations Management, 15*(1), 118–131.
Gupta, D., & Selvaraju, N. (2006). Performance evaluation and stock allocation in capacitated serial supply systems. *Manufacturing & Service Operations Management, 8*(2), 169–191.
Gustavsson, J., Cederberg, C., Sonesson, U., van Otterdijk, R., & Meybeck, A. (2011). *Global food losses and food waste*. Rome, Italy: The Food and Agriculture Organization of the United Nations.
Hanukov, G., Avinadav, T., Chernonog, T., Spiegel, U., & Yechiali, U. (2017). A queueing system with decomposed service and inventoried preliminary services. *Applied Mathematical Modelling, 47*, 276–293.
Hanukov, G., Avinadav, T., Chernonog, T., Spiegel, U., & Yechiali, U. (2018). Improving efficiency in service systems by performing and storing "preliminary services. *International Journal of Production Economics, 197*, 174–185.
Harchol Balter, M. (2013). *Performance modeling and design of computer systems*. Cambridge University Press.
Herbon, A. (2018). A non-cooperative game model for managing a multiple-aged expiring inventory under consumers' heterogeneity to price and time. *Applied Mathematical Modelling, 51*, 38–57.
Herbon, A., & Khmelnitsky, E. (2017). Optimal dynamic pricing and ordering of a perishable product under additive effects of price and time on demand. *European Journal of Operational Research, 260*(2), 546–556.
Hsieh, T. P., & Dye, C. Y. (2017). Optimal dynamic pricing for deteriorating items with reference price effects when inventories stimulate demand. *European Journal of Operational Research, 262*(1), 136–150.
Hsu, P. H., Wee, H. M., & Teng, H. M. (2010). Preservation technology investment for deteriorating inventory. *International Journal of Production Economics, 124*(2), 388–394.
Hu, P., Shum, S., & Yu, M. (2015). Joint inventory and markdown management for perishable goods with strategic consumer behavior. *Operations Research, 64*(1), 118–134.
Huang, J., Carmeli, B., & Mandelbaum, A. (2015). Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research, 63*(4), 892–908.
Hwang, J., Gao, L., & Jang, W. (2010). Joint demand and capacity management in a restaurant system. *European Journal of Operational Research, 207*(1), 465–472.
Jeganathan, K., Reiyas, M. A., Padmasekaran, S., & Lakshmanan, K. (2017). An $M/E_K/1/N$ queueing-inventory system with two service rates based on queue lengths. *International Journal of Applied and Computational Mathematics, 3*(1), 357–386.
Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM series on statistics and applied probability*. Philadelphia, PA: SIAM.
Krishnamoorthy, A., Manikandan, R., & Lakshmy, B. (2015). A revisit to queueing-inventory system with positive service time. *Annals of Operations Research, 233*(1), 221–236.

Please cite this article as: G. Hanukov et al., Performance improvement of a service system via stocking perishable preliminary services, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.10.027

Levy, Y., & Yechiali, U. (1975). Utilization of idle time in an M/G/1 queueing system. *Management Science, 22*, 202–211.
Levy, Y., & Yechiali, U. (1976). An M/M/s queue with servers' vacations. *INFOR, 14*, 153–163.
Li, Q., Yu, P., & X, Wu (2016). Managing perishable inventories in retailing: Replenishment, clearance sales, and segregation. *Operations Researc, 64*(6), 1270–1284.
Ma, Y., Liu, W. Q., & Li, J. H. (2013). Equilibrium balking behavior in the Geo/Geo/1 queueing system with multiple vacations. *Applied Mathematical Modelling, 37*(6), 3861–3878.
Mytalas, G. C., & Zazanis, M. A. (2015). An $M^X/G/1$ queueing system with disasters and repairs under a multiple adapted vacation policy. *Naval Research Logistics, 62*, 171–189.
Nair, A. N., Jacob, M. J., & Krishnamoorthy, A. (2015). The multi server $M/M/(s, S)$ queueing inventory system. *Annals of Operations Research, 233*(1), 321–333.
Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Baltimore, MD: Johns Hopkins University Press.
Pahl, J., & Voß, S. (2014). Integrating deterioration and lifetime constraints in production and supply chain planning: A survey. *European Journal of Operational Research, 238*(3), 654–674.
Perlman, Y., Elalouf, A., & Yechiali, U. (2018). Dynamic allocation of stochastically-arriving flexible resources to random streams of objects with application to kidney cross-transplantation. *European Journal of Operational Research, 265*(1), 169–177.
Rosenberg, E., & Yechiali, U. (1993). The $M^X/G/1$ queue with single and multiple vacations under the LIFO service regime. *Operations Research Letters, 14*(3), 171–179.
Shi, Y., & Lian, Z. (2016). Optimization and strategic behavior in a passenger–taxi service system. *European Journal of Operational Research, 249*(3), 1024–1032.
Van Do, T. (2015). A closed-form solution for a tollbooth tandem queue with two heterogeneous servers and exponential service times. *European Journal of Operational Research, 247*(2), 672–675.
Wang, W. C., Teng, J. T., & Lou, K. R. (2014). Seller's optimal credit period and cycle time in a supply chain for deteriorating items with maximum lifetime. *European Journal of Operational Research, 232*(2), 315–321.
Wang, J., Zhang, X., & Huang, P. (2017). Strategic behavior and social optimization in a constant retrial queue with the N-policy. *European Journal of Operational Research, 256*(3), 841–849.
Xu, Y., Scheller-Wolf, A., & Sycara, K. (2015). The benefit of introducing variability in single-server queues with application to quality-based service domains. *Operations Research, 63*(1), 233–246.
Yang, D. Y., & Wu, C. H. (2015). Cost-minimization analysis of a working vacation queue with N-policy and server breakdowns. *Computers & Industrial Engineering, 82*, 151–158.
Yang, C. T., Dye, C. Y., & Ding, J. F. (2015). Optimal dynamic trade credit and preservation technology allocation for a deteriorating inventory model. *Computers & Industrial Engineering, 87*, 356–369.
Yechiali, U. (2004). On the $M^X/G/1$ queue with a waiting server and vacations. *Sankhya, 66*, 159–174.
Zacharias, C., & Pinedo, M. (2017). Managing customer arrivals in service systems with multiple identical servers. *Manufacturing & Service Operations Management*.
Zhang, H., Shi, C., & Chao, X. (2016). Approximation algorithms for perishable inventory systems with setup costs. *Operations Research, 64*(2), 432–440.
Zhang, J., Wei, Q., Zhang, Q., & Tang, W. (2016). Pricing, service and preservation technology investments policy for deteriorating items under common resource constraints. *Computers & Industrial Engineering, 95*, 1–9.
Zhao, N., & Lin, Z. T. (2011). A queueing-inventory system with two classes of customers. *International Journal of Production Economics, 129*, 225–231.
Zhou, W., Huang, W., & Zhang, R. (2014). A two-stage queueing network on form postponement supply chain with correlated demands. *Applied Mathematical Modelling, 38*(11), 2734–2743.

Please cite this article as: G. Hanukov et al., Performance improvement of a service system via stocking perishable preliminary services, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.10.027