International Journal of Production Research

ISSN: 0020-7543 (Print) 1366-588X (Online) Journal homepage: https://www.tandfonline.com/loi/tprs20

A multi-server system with inventory of preliminary services and stock-dependent demand

Gabi Hanukov, Tal Avinadav, Tatyana Chernonog & Uri Yechiali

To cite this article: Gabi Hanukov, Tal Avinadav, Tatyana Chernonog & Uri Yechiali (2020): A multi-server system with inventory of preliminary services and stock-dependent demand, International Journal of Production Research, DOI: 10.1080/00207543.2020.1762945

To link to this article: https://doi.org/10.1080/00207543.2020.1762945

Published online: 14 May 2020.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tprs20

International Journal of Production Research, 2020
https://doi.org/10.1080/00207543.2020.1762945

A multi-server system with inventory of preliminary services and stock-dependent demand

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

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

(*Received 17 December 2019; accepted 23 April 2020*)

This study is motivated by industries in which products can be partially prepared and stored before demand occurs, while demand is stock-dependent. We study a multi-server system in which the servers utilise their idle time to produce and store 'preliminary services' (PSs) in order to reduce customers' sojourn time and, as a result, stimulating demand by creating the anticipation of a shorter sojourn time. In order to facilitate closed-form solutions, we analyse a Markovian queueing-inventory model and apply matrix geometric (MG) methods. In contrast to most applications in which the rate matrix R of the MG analysis is calculated numerically, our analysis enables derivation of explicit solutions for the entries of R and discovers their relation to Catalan numbers, allowing a rapid solution for large systems. Consequently, the system's stability condition is readily obtained and shown to be identical to that of a regular M/M/s queue. Two models are developed: one for non-perishable PSs, the other for perishable ones. An economic analysis is provided for two case studies: a bike store and a pizza store. We observe that the reward has low sensitivity to deviation from the optimal PSs capacity and high sensitivity to deviation from the optimal promotional level.

Keywords: Markov process; queueing-inventory system; stock-dependent demand; matrix geometric analysis; economic analysis

1. Introduction

Reducing customer waiting times in service systems is a common objective, which aims to reduce costs associated with customers' sojourn time in the system and increase their satisfaction. Usually, it is achieved by using faster servers (e.g. Hwang, Gao, and Jang 2010; Guo and Zhang 2013) or by hiring additional servers. However, these methods are costly. Recent studies (Hanukov et al. 2017, 2018, 2019a, 2019b) have suggested using a server's idle time to produce and store preliminary services, which will be used to serve future customers more quickly and thus increase the productivity of the service system without investing in additional or better resources. This innovative approach is modelled using a queueing-inventory system represented by a two-dimensional Markov process (see, e.g. Cox and Miler 1965), which is solved using matrix geometric analysis. The main question in these models is: How many preliminary services should be produced and stored when the server is idle? Actually, this approach represents an interesting contrast with another, similar scenario that has been extensively studied in the literature whereby servers' idle times may be used for carrying out ancillary duties (also known as 'vacation models'; see, for example, Levy and Yechiali 1975, 1976; Rosenberg and Yechiali 1993; Boxma, Schlegel, and Yechiali 2002; Yechiali 2004; Baek et al. 2014; Yang and Wu 2015; Mytalas and Zazanis 2015; Guha, Goswami, and Banik 2016; Barron 2018), which may increase customer waiting times. Thus far, the innovative approach of using idle times to perform preliminary services has been applied (analytically) only to a single-server service system and a homogeneous Poisson (customer) arrival process.

The current work extends this innovative approach (Hanukov et al. 2017, 2018, 2019a, 2019b) to the framework of a multi-server service system with a stock-dependent customer arrival rate. Multi-server systems are widespread in real-life businesses, but in the case of queueing-inventory systems, are considerably more complex, and thus are more difficult to analyse. As for stock-dependent demand, it is known (see, e.g. Urban 2005) that customers buy more units of a product when the level of the presented inventory is higher, which is due to marketing effects (e.g. a wider selection). In our model, the stock of preliminary services has an additional positive effect on customer demand, since it implies a shorter customer sojourn time in the system. To model this effect, we use a state-dependent customer arrival rate (Gupta 1967; Winston 1978; Guo and Hassin 2011; Zhao and Lian 2011). In addition, we consider the deterioration of preliminary services

*Corresponding author. Email: german.kanukov@biu.ac.il

© 2020 Informa UK Limited, trading as Taylor & Francis Group

while being held in stock. In particular, we assume an exponential lifetime of preliminary services, where spoiled units are discarded.

Our model is mostly motivated by the fast food industry, in which food such as hamburger patties or basic pizzas can be prepared before demand occurs, and only upon the arrival of a customer are they heated up, adjusted according to the specific customer requirements, and then served. Another domain in which our model is applicable is bicycle retailing, where mechanics can assemble parts of a bicycle before a customer's arrival, and subsequently assemble the remaining parts in accordance with the customer's specific requirements and preferences. Handmade nameplates for doors are another example in which service can be split up. The server can produce basic nameplates from wood, ceramic or glass before an order is made, and complete the nameplate upon request (e.g. by writing the name or adding some form of decoration).

The main contributions of this paper are:

- This is the first study to investigate both of the aforementioned positive effects of preliminary services (i.e. utilising the server idle time and stimulating demand) on the service system performance.
- Closed-form expressions for the steady-state probabilities of the system states are derived, leading to the calculation of various performance measures for large problems in a relatively short time-frame.
- The relation of these expressions to Catalan numbers is shown. This result is notable due to the fact that, in typical applications of matrix geometric analysis, explicit calculation of the entries of the rate matrix R is rarely possible.
- It is proved analytically that the stability condition of our model is identical to that of a standard M/M/2 queue.

2. Literature review

Prior studies that have considered queueing-inventory systems include the following: Zhao and Lian (2011) studied a service facility in which each service uses one item in the associated inventory, which is supplied with exponentially distributed lead time. Wang et al. (2013) modelled a simple production-inventory-queue system and formulated the optimal integrated control problem as a continuous Markov decision process. Adacher and Cassandras (2014) utilised a queueing model to tackle the lot-sizing problem in manufacturing systems. Avşar and Zijm (2014) presented approximate queueing models for capacitated multi-stage inventory systems under base-stock control. Altendorfer and Minner (2015) modelled a production system as an M/M/1 queue with input rates that are dependent on queue length and random customer-required lead-time. They then developed a heuristic solution for the optimal capacity investment problem. Chebolu-Subramanian and Gaukler (2015) used queueing theory to analyse product contamination in a multi-stage food supply chain with inventory. 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 a certain probability to a customer at his service completion epoch. The authors investigated optimisation problems associated with both models. Nair, Jacob, and Krishnamoorthy (2015) considered a multi-server Markovian queueing 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. Wang, Lan, and Jiang (2016) explored the impact of customer impatience on the performance of a system that consists of make-to-stock queueing and an M/M/1 service-orders-processed subsystem. Otten, Krenzle, and Daduna (2016) used queueing and inventory theories to analyse a two-echelon production system with a central supplier who delivers raw material to servers at several locations, each with a local inventory. Wang and Zhang (2017) investigated the individual equilibrium strategy and the socially optimal strategy, along with the optimal price, in a single-server service-inventory system while considering customer-choice behaviour.

Inventory control of perishable items has been thoroughly investigated in the Operations Management literature. These studies deal with order quantity at fixed intervals or according to inventory position. For example, Avinadav and Arponen (2009) extended the classical EOQ model to products with a fixed expiry date and a declining demand rate due to a reduction in the quality of the product with time. Hu, Shum, and Yu (2015) formulated and analysed a dynamic model of inventory and pricing decisions for perishable goods under uncertainty, where every period consists of two phases: a clearance phase and a regular-sales phase. Chao et al. (2015) developed approximation algorithms with worst-case performance guarantees for stochastic periodic-review perishable inventory systems. They considered both backlogging and lost-sales models. Chernonog (2020) and Avinadav (2020) investigated a two-echelon supply chain of a perishable product in which a manufacturer and a retailer interact via a wholesale price contract and via a revenue sharing contract, respectively. For further discussion on this topic, see for example, (Cooper 2001; Berk and Gürler 2008; Avinadav, Herbon, and Spiegel 2013, 2014; Chen, Pang, and Pan 2014; Li, Yu, and Wu 2016; Zhang, Shi, and Chao 2016; Avinadav et al. 2017; Chernonog and Avinadav 2017; Herbon 2017, 2018; Herbon and Khmelnitsky 2017; Dolgui et al. 2018; Lacomme et al. 2018; Abouee-Mehrizi et al. 2019; Barron 2019; Barron and Baron 2020). In our study, we consider the less explored scenario in which inventory is accumulated only when the server is idle (a stochastic process) and via production (also a stochastic process). Inventory is depleted either via demand or due to spoilage of pre-prepared food units (both are stochastic processes).

Relatively few papers have combined queueing systems and inventories of perishable items. Ioannidis et al. (2013) considered problems of inventory control for make-to-stock production systems with perishable inventory and queueing of impatient customers. Jeganathan et al. (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. Kırcı, Biçer, and Seifert (2019) considered a two-echelon supply chain for a perishable product where the production process is modelled using an M/M/1 queueing system and the retailing process includes inventory decisions. However, none of the above papers considered utilising the server's idle time to produce preliminary services (PSs).

Utilising a server's idle time to produce preliminary services has been recently investigated in the Operations Management literature. Hanukov et al. (2017) studied a queueing system with decomposed service in which they focused on analysing the distribution of customer waiting times. Hanukov et al. (2018) analysed the effect of the preliminary services' capacity on the server's reward, where the service can be conducted either as one continuous process or as a two-phase process. Whereas the latter two papers do not consider perishability of the preliminary services, Hanukov et al. (2019a) analysed a fast food service queueing-inventory system, which includes perishability considerations. Hanukov et al. (2020) extended the latter study to include two types of customer: fastidious and strategic. While all the above papers considered a single-server system with a fixed customer arrival rate, only Hanukov et al. (2019b) and the current study consider multi-server queueing, which is more common in the fast food industry, and analyses the use of preliminary services as a means of increasing demand. This paper extends Hanukov et al. (2019b) by considering the perishability effect and by providing two case studies to show the applicability of this model to real-life problems (specifically, a bike store and a pizza store), which are followed by comprehensive sensitivity analyses.

3. Model formulation

Assume a service system with two servers, in which full-service (FS) time is exponentially-distributed with mean $1/\mu$, and customer arrival follows a Poisson process with rate λ. The service can be split into two phases: a preliminary service (PS), which can be carried out in the absence of the customer, and a complementary service (CS), which can be provided only when the customer is present. When a server is idle, he/she can produce preliminary services. In line with prior studies that involve make-to-stock queueing (Benjaafar, Cooper, and Mardan 2011; Flapper, Gayon, and Vercraene 2012; Iravani, Kolfal, and Van Oyen 2011; Hanukov et al. 2017), the preparation time of the PS is assumed to be exponentially distributed with parameter α. The PSs are stored until the arrival of customers, and are used to reduce the sojourn time of future customers. The total number of preliminary services in stock is limited to n (capacity), and when the inventoried PSs reach this capacity, the servers stop producing them and resume an idle position. When a customer reaches the front of the queue and a preliminary service is available, his/her CS phase starts immediately. The CS time is exponentially distributed with mean $1/\beta$, where $\beta > \mu$. That is, the CS duration is stochastically shorter than the FS duration (first-order stochastic dominance). When PSs are available, the customer arrival rate increases to δ $(> \lambda)$ until the number of customers is equal to the number of PSs (both of which are observable by the customers).

The process can be formulated as a three-dimensional continuous-time Markov chain with a state space $\{L_t, S_t, U_t\}$, where $L_t \in \{0, 1, 2, \ldots, \infty\}$ denotes the number of customers in the system at time t, $S_t \in \{0, 1, 2, \ldots n\}$ denotes the total number of PSs at time t, and $U_t \in \{0, 1, 2\}$ denotes the number of customers in their CS phase at time t. Let $L \equiv \lim_{t\to\infty} L_t$, $S \equiv \lim_{t\to\infty} S_t$ and $U \equiv \lim_{t\to\infty} U_t$. Let $p_{i,j,k} = \Pr(L = i, S = j, U = k)$ denote the joint probability distribution of the latter three-dimensional process. Thus $p_{i,j,k}$ represents the long-run fraction of time that the system stays in state $(L = i, S = j, U = k)$. As an illustration, Figure 1 below depicts the system's states and its corresponding transition-rate diagram for the case where the PS capacity is $n = 3$. Each circle in the diagram depicts a state $(L = i, S = j, U = k)$ and the arrows indicate the transition directions and rates.

Our goals are (i) to obtain the system's stability condition (i.e. the condition on the parameters that will ensure that the queue size does not grow beyond any given bound) and (ii) to derive the system's steady-state probabilities (i.e. $p_{i,j,k}$ when the system is stable). Subsequently, various performance measures, such as mean queue size, mean sojourn time, mean PS inventory level, and mean time a PS spends in inventory, are calculated. For these purposes, matrix geometric analysis (Neuts 1981) is applied, which entails calculation of the so-called rate matrix R (see below). The system's states are arranged in the following lexicographic order ($i = 2, 3, \ldots$):

$$\{(0,0,0), (0,1,0), \ldots, (0,n,0); (1,0,0), (1,1,1), (1,1,0), \ldots, (1,n,1), (1,n,0); \ldots; \\ (i,0,0), (i,1,1), (i,2,2), (i,2,1), \ldots, (i,n,2), (i,n,1); \ldots\},$$

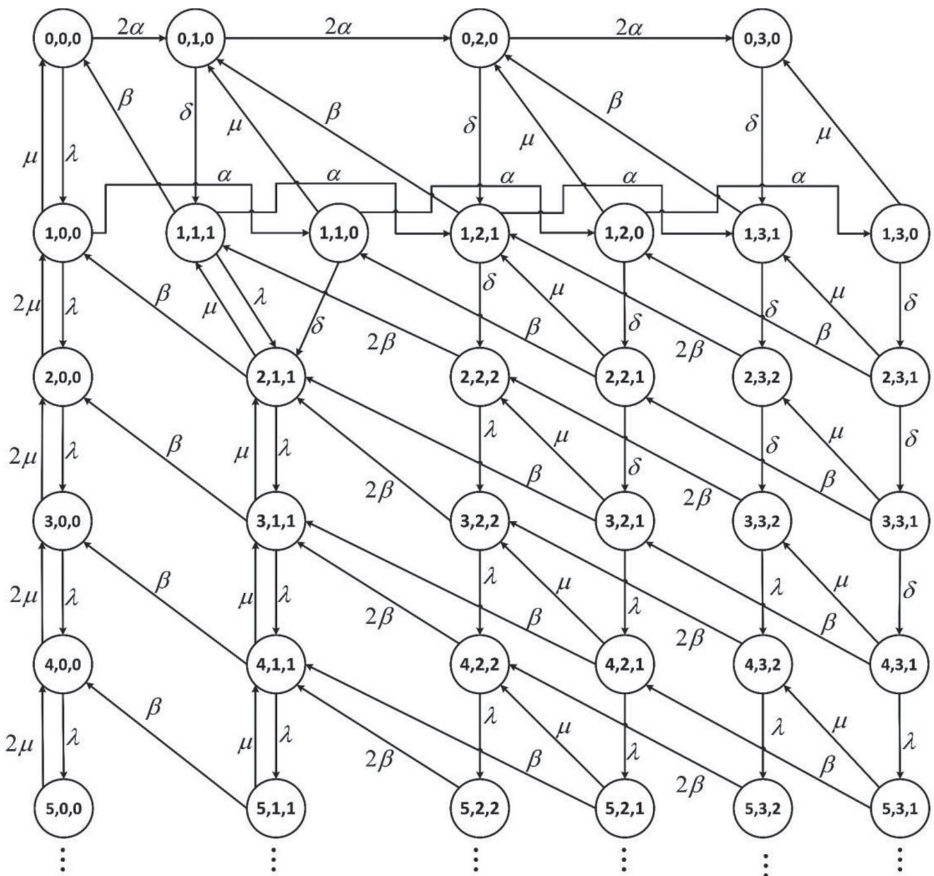

Figure 1. System states and transition-rate diagram for $n = 3$.

which allows us to construct the system's infinitesimal generator matrix, denoted by Q:

$$Q = \begin{pmatrix} B_{1,1} & B_{1,2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ B_{2,1} & B_{2,2} & B_{2,3} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & B_{3,1} & B_{3,2} & B_{3,3} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & A_2 & B_{4,2} & B_{4,3} & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \ddots & & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & A_2 & B_{n+1,2} & B_{n+1,3} & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1}$$

where matrices $B_{i,j}$ are given in Appendix 1, and matrices $A_0 \equiv [a_0^{v,t}]$, $A_1 \equiv [a_1^{v,t}]$ and $A_2 \equiv [a_2^{v,t}]$ are given by $A_0 = \lambda I_{2n\times 2n}$,

$$a_1^{v,t} = \begin{cases} -(\lambda + 2\mu) & v = t = 1 \\ -(\lambda + \mu + \beta) & v = t = 2, 4, 6, 8, \ldots, 2n \\ -(\lambda + 2\beta) & v = t = 3, 5, 7, \ldots, 2n-1 \\ 0 & \text{Otherwise} \end{cases}, \tag{2}$$

$$a_2^{v,t} = \begin{cases} 2\mu & v = t = 1 \\ \mu & v = t = 2 \text{ and } v = t+1 = 4, 6, 8, \ldots, 2n \\ \beta & v = t+1 = 2 \text{ and } v = t+2 = 4, 6, 8, \ldots, 2n \\ 2\beta & v = t+1 = 3 \text{ and } v = t+2 = 5, 7, 9, \ldots, 2n-1 \\ 0 & \text{Otherwise} \end{cases}. \tag{3}$$

4. Steady-state analysis

We now claim that the condition for the stability of the system is identical to that of a regular M/M/2 queue without preliminary services (i.e. a Markovian service system with two parallel identical servers).

THEOREM 4.1 *The stability condition of the queueing-inventory system is $\lambda < 2\mu$.*

Proof According to Hanukov and Yechiali (2020), when each of the matrices A_0, A_1 and A_2 is lower triangular, which is the case in our model, the stability condition is given by $a_0^{1,1} < a_2^{1,1}$. By substituting $a_0^{1,1} = \lambda$ and $a_2^{1,1} = 2\mu$ the claim is proved. ■

Theorem 4.1 implies that although preparing preliminary services improves the system's performance, it does not alter its stability condition. This result can be further explained as follows: due to the stochastic nature of the system, at some point in time, both servers will be fully occupied and will not be able to produce any PSs, so the system will become a regular M/M/2 queue. This observation leads us to state the following:

Stability: The system stability condition can be readily extended to the case where the underlying process is a regular M/M/*s* queue, namely, $\lambda < s\mu$.

The matrix geometric analysis is based on derivation of a so-called rate matrix $R \equiv [r^{v,t}]_{2n\times 2n}$, which is obtained by solving the quadratic matrix equation (see Neuts 1981).

$$A_0 + RA_1 + R^2 A_2 = 0_{2n\times 2n}. \tag{4}$$

Let us define the $(n+1)$-dimensional vector of probabilities

$$\vec{p}_0 \equiv \begin{pmatrix} p_{0,0,0} & p_{0,1,0} & p_{0,2,0} & p_{0,3,0} & \cdots & p_{0,n,0} \end{pmatrix},$$

the $(2n+1)$-dimensional vector of probabilities

$$\vec{p}_1 \equiv \begin{pmatrix} p_{1,0,0} & p_{1,1,1} & p_{1,1,0} & p_{1,2,1} & p_{1,2,0} & p_{1,3,1} & p_{1,3,0} & \cdots & p_{1,n,1} & p_{1,n,0} \end{pmatrix}$$

and the $(2n)$-dimensional vectors of probabilities

$$\vec{p}_i \equiv \begin{pmatrix} p_{i,0,0} & p_{i,1,1} & p_{i,2,2} & p_{i,2,1} & p_{i,3,2} & p_{i,3,1} & \cdots & p_{i,n,2} & p_{i,n,1} \end{pmatrix} \quad i = 2,3,4,\ldots.$$

Following Neuts (1981), the vectors $\vec{p}_i$ satisfy

$$\vec{p}_i = \vec{p}_{n+1} R^{i-(n+1)}, \quad i = n+1, n+2, n+3, \ldots, \infty. \tag{5}$$

In most cases, the entries $r^{v,t}$ of the matrix R are found by numerical calculations (see Chapter 8 in Latouche and Ramaswami 1999), mostly using *successive substitutions* (Neuts 1981, 37). However, for some special cases, the matrix can be derived analytically. The matrices A_0, A_1 and A_2 are lower triangular, and thus we can apply Theorem 4.1 of Hanukov and Yechiali (2020) to provide closed-form formulae for the elements of the matrix R; the results are detailed in Appendix 2. Since, in the current model, the matrices A_0, A_1 and A_2 are indeed lower triangular, it is possible to use these formulae to derive closed-form expressions for all $r^{v,t}$, as given in Theorem 4.2.

Let C_m be the m-th Catalan number (i.e. $C_m \equiv ((2m)!/(m+1)!m!)$, see Koshy 2008), where m is a non-negative integer. Then,

THEOREM 4.2 *The entries of matrix R are:*

(i) $r^{v,t} = 0$ *for* $v < t$, *(ii)* $r^{1,1} = \frac{\lambda}{2\mu}$, *(iii)* $r^{2,2} = \frac{\beta+\mu+\lambda-\sqrt{(\beta+\mu+\lambda)^2-4\lambda\mu}}{2\mu}$,

(ii) $r^{v,1} = \frac{2\mu\sum_{\tau=2}^{v-1} r^{v,\tau} r^{\tau,1} + \beta \sum_{\tau=2}^{v} r^{v,\tau} r^{\tau,2}}{2\mu(1-r^{v,v})}$, $v = 2,3,4,\ldots,2n,$

(iii) $r^{v,2} = \frac{\beta\left(\sum_{\tau=3}^{v-1} r^{v,\tau} r^{\tau,2} + 2\sum_{\tau=3}^{v} r^{v,\tau} r^{\tau,3} + \sum_{\tau=4}^{v} r^{v,\tau} r^{\tau,4}\right)}{\beta+\lambda+\mu(1-r^{2,2}+r^{v,v})}$, $v = 3,4,5,\ldots,2n,$

(iv) $r^{v,t} = \begin{cases} 0 & v = t+1, t+3, \ldots, 2n-1 \\ \frac{C_{0.5(v-t)}\beta^{0.5(v-t)}\lambda^{0.5(v-t+2)}}{(\lambda+\mu+\beta)^{v-t+1}} & v = t, t+2, \ldots, 2n \end{cases}$, $t = 4,6,8,\ldots,2n,$

(v) $r^{v,t} = \begin{cases} \frac{\mu\sum_{\tau=t+1}^{v} r^{v,\tau} r^{\tau,t+1} + 2\beta\sum_{\tau=t+2}^{v} r^{v,\tau} r^{\tau,t+2}}{2\beta+\lambda} & v = t+1, t+3, \ldots, 2n \\ \frac{2^{0.5(v-t)} C_{0.5(v-t)}\beta^{0.5(v-t)}\lambda^{0.5(v-t+2)}}{(\lambda+2\beta)^{v-t+1}} & v = t, t+2, \ldots, 2n-1 \end{cases}$, $t = 3,5,7,\ldots,2n-1.$

Proof (i), (ii), (iii), (iv), (v) and the first term of (vii) are derived directly from the formulae given in Appendix 2. The proofs for (vi) and for the second term of (vii) are given in Appendix 3. ■

Let $\vec{p} \equiv (\vec{p}_0, \vec{p}_1, \vec{p}_2, \ldots)$. Then, to calculate $\vec{p}_i$, $i = 0, 1, \ldots, n+1$, we use a subset of equations from $\vec{p}Q = 0$ combined with the normalisation equation $\vec{p}_0 \cdot \vec{e} + \vec{p}_1 \cdot \vec{e} + \sum_{i=2}^{\infty} \vec{p}_i \cdot \vec{e} = 1$, where $\vec{e}$ is a vector whose elements are all equal to 1.

Consequently, we solve the following linear set:

$$\begin{array}{ll} \vec{p}_0 B_{1,1} + \vec{p}_1 B_{2,1} = \vec{0} & \\ \vec{p}_0 B_{1,2} + \vec{p}_1 B_{2,2} + \vec{p}_2 B_{3,1} = \vec{0} & \\ \vec{p}_{i-1} B_{i,3} + \vec{p}_i B_{i+1,2} + \vec{p}_{i+1} A_2 = \vec{0} & i = 2, 3, \ldots, n \\ \vec{p}_n B_{n+1,3} + \vec{p}_{n+1}(A_1 + RA_2) = \vec{0} & \\ \vec{p}_0 \cdot \vec{e} + \vec{p}_1 \cdot \vec{e} + \sum_{i=2}^{n} \vec{p}_i \cdot \vec{e} + \vec{p}_{n+1}[I-R]^{-1} \cdot \vec{e} = 1. & \end{array} \tag{6}$$

5. Performance measures

Based on the steady-state probabilities, we show how to calculate relevant system performance measures. For a given capacity of preliminary services, n, let $L(n)$ and $L_q(n)$ be the mean number of customers in the system and in the queue, respectively. Similarly, let: $W(n)$ and $W_q(n)$ be the mean sojourn time of a customer in the system and in the queue, respectively; $S(n)$ and $S_q(n)$ be the mean number of PSs in the system and in the inventory, respectively; and $T(n)$ and $T_q(n)$ be the mean time a PS resides in the system and in the inventory, respectively.

Using (5) and the equations $\sum_{i=0}^{\infty} R^i = [I-R]^{-1}$ and $\sum_{i=0}^{\infty}(i+1)R^i = [I-R]^{-2}$, we obtain:

$$L(n) = \vec{p}_1\vec{e} + \sum_{i=2}^{n} i\vec{p}_i\vec{e} + \vec{p}_{n+1}(n[I-R]^{-1} + [I-R]^{-2})\vec{e}. \tag{7}$$

$$L_q(n) = L(n) - 2(1 - \vec{p}_0\vec{e}) + \vec{p}_1\vec{e}. \tag{8}$$

In order to calculate the mean sojourn time of customers in the system and in the queue, we first have to calculate the effective customer arrival rate λ_{eff}. Let p_δ be the proportion of time for which customers arrive with rate δ, and $p_\lambda = 1 - p_\delta$ the proportion of time for which customers arrive with rate λ. Then, the effective customer arrival rate is obtained as follows

$$\lambda_{eff} = \delta p_\delta + \lambda p_\lambda = (\delta - \lambda)p_\delta + \lambda, \tag{9}$$

where

$$p_\delta = \sum_{i=0}^{1}\sum_{j=1}^{n} p_{i,j,0} + \sum_{j=2}^{n} p_{1,j,1} + \sum_{i=2}^{n}\sum_{j=i}^{n} p_{i,j,1} + \sum_{i=2}^{n-1}\sum_{j=i+1}^{n} p_{i,j,2}. \tag{10}$$

Thus, $W(n) = L(n)/\lambda_{eff}$ and $W_q(n) = L_q(n)/\lambda_{eff}$.

In order to obtain $S(n)$, we define three column vectors $\vec{v}_{n+1} = \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots & n \end{pmatrix}^T$, $\vec{v}_{2n+1} = (0 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3 \quad 3 \quad \cdots \quad n \quad n)^T$ and $\vec{v}_{2n} = \begin{pmatrix} 0 & 1 & 2 & 2 & 3 & 3 & \cdots & n & n \end{pmatrix}^T$. Thus, by using Equation (4) and algebraic manipulations, we get

$$S(n) = \vec{p}_0\vec{v}_{n+1} + \vec{p}_1\vec{v}_{2n+1} + \sum_{i=2}^{n} \vec{p}_i\vec{v}_{2n} + \vec{p}_{n+1}[I-R]^{-1}\vec{v}_{2n}. \tag{11}$$

Similarly, in order to obtain $S_q(n)$, we define three column vectors $\vec{u}_{n+1} = \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots & n \end{pmatrix}^T$, $\vec{u}_{2n+1} = \begin{pmatrix} 0 & 0 & 1 & 1 & 2 & 2 & 3 & \cdots & n-1 & n \end{pmatrix}^T$ and $\vec{u}_{2n} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 2 & 2 & 3 & \cdots & n-2 & n-1 \end{pmatrix}^T$, so that

$$S_q(n) = \vec{p}_0\vec{u}_{n+1} + \vec{p}_1\vec{u}_{2n+1} + \sum_{i=2}^{n} \vec{p}_i\vec{u}_{2n} + \vec{p}_{n+1}[I-R]^{-1}\vec{u}_{2n}. \tag{12}$$

In the case where the inventory level is less than n, preliminary services are produced by both servers when there are no customers in the system, and by one of the servers when there is only one customer in the system. Therefore, the effective production rate of PSs is

$$\alpha_{eff} = 2\alpha(\vec{p}_0\vec{e} - p_{0,n,0}) + \alpha(\vec{p}_1\vec{e} - p_{1,n,0} - p_{1,n,1}). \tag{13}$$

Consequently, the mean periods of time for which a PS resides in the system and in the inventory, respectively, as obtained from Little's law, are as follows: $T(n) = S(n)/\alpha_{eff}$ and $T_q(n) = S_q(n)/\alpha_{eff}$.

6. Economic analysis

In this section, we provide an economic analysis of our queueing-inventory model. Let REV be the net revenue from supplying a service to a customer, so that the expected system's revenue rate is equal to $REV \cdot \lambda_{eff}$. Let c be the sojourn cost per unit time per customer in the system, and let h be the capacity cost per unit time. Assume that the server can publicise the existence of available preliminary services, which stimulates demand (due to a shorter sojourn time). The publicity may be carried out at different levels. For example, the server could rent screen time for promotion at different locations near its store, where each screen may increase customers' arrival rate. Another possibility would be to transmit promotional messages to customers' smartphones, which could take place over a larger distance to reach more customers. Thus, we denote by $A(\delta)$ the cost rate of making the number of PSs observable (hereafter transparency cost) as a function of the desired arrival rate δ. We assume that the transparency cost is a convex increasing function of the gap between the desired and the actual arrival rates. Specifically, we use $A(\delta) = \psi(\delta - \lambda)^\tau p_\delta$,where ψ (> 0) is a scale coefficient, τ (> 1) is a shape coefficient, and p_δ is defined in Equation (10). The objective is to maximise the total expected profit per unit time, Z, by controlling the PS capacity n and the desired arrival rate when PSs are available, δ. We add the arguments n and δ to each term that includes them, such that the optimisation problem is defined as follows:

$$\max_{\substack{n\in\{0,1,2\ldots\} \\ \delta\geq\lambda}} \{Z(n,\delta) = REV \cdot \lambda_{eff}(n,\delta) - A(n,\delta) - cL(n,\delta) - hn\}. \tag{14}$$

To illustrate the behaviour of the expected profit as a function of the PS capacity n and the desired arrival rate when PSs are available δ, we use a numerical example of a bike store in Israel. The estimation is that each year 200,000 bikes are sold[1] via 300 bike stores,[2] which are open between 200 and –250 days per year. Therefore, we use $\lambda = 3[\frac{\text{customers}}{\text{day}}]$. It takes about two hours to build a complete bike from its components and then tune it according to the customer requirements, so, assuming 8 working hours per day, we use a full-service rate of $\mu = 4[\frac{\text{customers}}{\text{day}}]$. We further assume that it takes about an hour[3] to assemble a bike out of a box without the presence of a customer, while a similar length of time is required to tune the bike according to the customer requirements (explanations, upgrades, oiling, testing, etc.); thus we use $\alpha = 7[\frac{\text{PSs}}{\text{day}}]$ and $\beta = 8[\frac{\text{CSs}}{\text{day}}]$. The selling price of bikes varies from dozens of dollars (for children's bikes) to hundreds of dollars or more (for professional bikes),[4] so we use an estimate of the net revenue (which is about one third of the selling price (see note 1)) of $REV = 300[\frac{\$}{\text{customers}}]$. We assume that customers who are waiting for their bike to be assembled and tuned cost the store owner about \$6 per hour; so we use $c = 50[\frac{\$}{\text{day}\times\text{customer}}]$. The cost of renting a store is about \$50 per square meter per month (see note 1), and since an assembled bike occupies approximately 1 square meter, we use $h = 1.5[\frac{\$}{\text{day}\times\text{PS}}]$. In addition, we use $\psi = 100\left[\frac{\$\times(\text{day})^{\tau-1}}{(\text{customer})^\tau}\right]$ and $\tau = 1.6$, implying that the promotion required to attract each buyer costs, per day, at least one third of the net revenue from selling a bike. The results are given in Table 1 and Figure 2.

Table 1 shows that the optimal PS capacity and enhanced customer arrival rate are $n^* = 14$ and $\delta^* = 6$, respectively, and that the point corresponding to these values is the unique optimal solution of the expected profit maximisation problem over the domain $n \leq 20$, $3 \leq \delta \leq 11$. One insight from this case study is that the reward is less sensitive to the capacity of the pre-prepared bikes in the neighbourhood of its optimal value (i.e. there is a change of no more than 1.3% within the range $6 \leq n \leq 19$), so the optimal value from the model can be used as is. In contrast, the reward is more sensitive to deviation from δ^*. Therefore, since $A(\delta)$ is taken as a proxy for the actual promotion cost, we recommend for the store owner to use a learning strategy by making small changes in the promotion expenditure until the best policy is determined.

6.1. Sensitivity analysis

Now we investigate the effect of the monetary parameters (REV, c, h, ψ) on the decision variables (n, δ) and on the associated profit (Z). In addition, we calculate the ratio between the profit when using preliminary services and that of a standard M/M/2 system: $\xi = Z(n^*, \delta^*)/Z_{M/M/2}$. The results are presented in Figures 3–10. Figure 3 shows that both the optimal PS capacity (n^*) and the optimal investment in publishing the existence of available PSs (which is correlated with δ^*)

Table 1. The expected profit for various values of n and δ, where the maximum expected profit is shown in bold typeface.

$n\backslash\delta$	3	4	5	6	7	8	9	10	11
0	866.5	958.5	984.6	976.8	948.8	908.1	859.2	804.9	747.0
1	870.4	1009.3	1052.6	1042.6	999.7	936.1	859.5	774.8	685.3
2	871.1	1031.7	1084.8	1074.2	1023.4	948.0	857.8	759.2	656.3
3	870.6	1043.0	1103.2	1093.0	1037.5	954.4	855.6	748.6	638.3
4	869.5	1048.9	1114.6	1105.4	1046.8	958.2	853.3	740.7	625.7
5	868.3	1051.9	1121.9	1114.1	1053.3	960.5	851.0	734.4	616.3
6	866.9	1053.2	1126.7	1120.4	1058.1	961.9	848.8	729.2	609.0
7	865.5	1053.5	1129.8	1125.0	1061.6	962.7	846.6	724.7	603.1
8	864.0	1053.2	1131.7	1128.3	1064.3	963.1	844.5	720.8	598.2
9	862.5	1052.4	1132.7	1130.8	1066.3	963.1	842.5	717.4	594.2
10	861.0	1051.4	1133.1	1132.5	1067.7	962.9	840.5	714.3	590.7
11	859.5	1050.2	1133.1	1133.7	1068.8	962.5	838.6	711.6	587.7
12	858.0	1049.0	1132.7	1134.5	1069.5	962.0	836.8	709.0	585.0
13	856.5	1047.6	1132.0	1134.9	1070.0	961.3	834.9	706.7	582.5
14	855.0	1046.2	1131.2	**1135.0**	1070.2	960.6	833.1	704.4	580.3
15	853.5	1044.8	1130.2	1134.9	1070.3	959.7	831.4	702.3	578.2
16	852.0	1043.3	1129.1	1134.5	1070.2	958.8	829.7	700.4	576.3
17	850.5	1041.8	1127.9	1134.1	1069.9	957.8	828.0	698.5	574.4
18	849.0	1040.3	1126.6	1133.4	1069.6	956.8	826.3	696.6	572.7
19	847.5	1038.9	1125.3	1132.7	1069.1	955.7	824.6	694.8	571.0
20	866.5	958.5	984.6	976.8	948.8	908.1	859.2	804.9	747.0

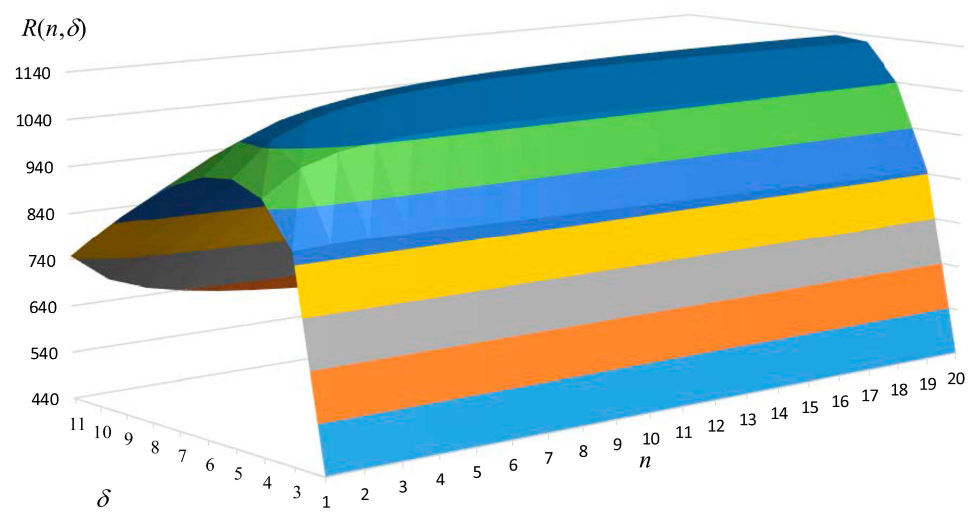

Figure 2. The expected profit as a function of n and δ.

have a positive trend when the revenue from supplying a bike to a customer (REV) increases. This is because a higher payoff per customer motivates the server to increase the effective arrival rate (λ_{eff}), either by publishing the existence of preliminary services or by increasing their capacity. Figure 4 indicates that, as predicted, the total expected profit both in the proposed system (Z) and in the classical M/M/2 system ($Z_{M/M/2}$) increases with REV, as do the profit ratio (ξ) and the absolute improvement in profits due to using preliminary services (i.e. the difference in height between the blue and the grey bars). Figure 5 shows that δ^* is robust to the value of h, because the latter represents the capacity cost and not the inventory cost, whereas, as expected, n^* decreases with h. Figure 6 shows that both Z and ξ decrease with h because increasing the capacity cost reduces the advantage of providing preliminary services, and thus diminishes the profitability of such a system. Figure 7 illustrates that both δ^* and n^* decrease as the customers' sojourn cost c increases, due to the reduced motivation to encourage arrivals when the sojourn time is more expensive. Figure 8 shows that, as expected, Z decreases with c; however, ξ increases with c, which implies that PSs play a more significant role in accelerating the service as the sojourn cost per unit time increases. Finally, Figures 9 and 10 demonstrate that, as expected, n^*, δ^*, Z and ξ all decrease with ψ.

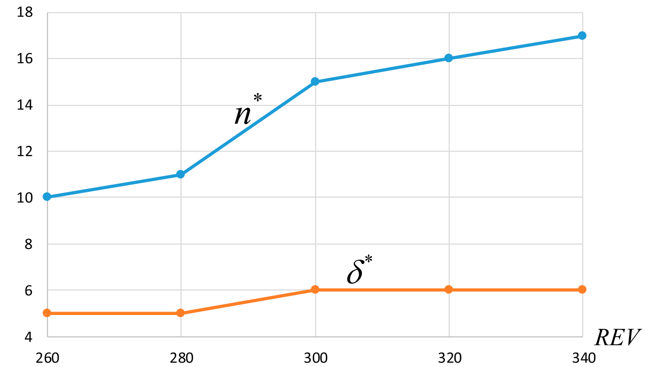

Figure 3. Optimal n and δ as a function of REV.

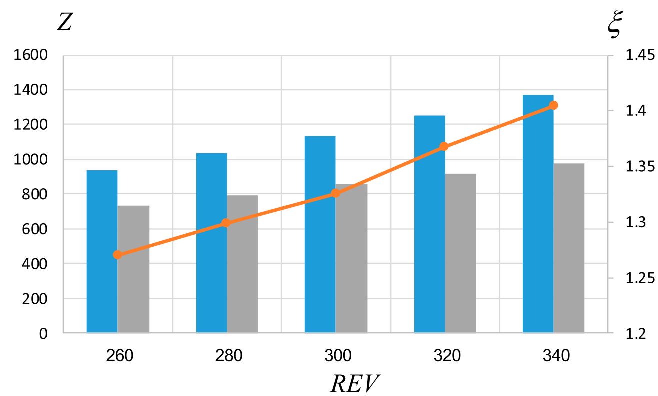

Figure 4. $Z(n^*, \delta^*)$ (left bars), $Z_{M/M/2}$ (right bars), and ξ (curve) as a function of REV.

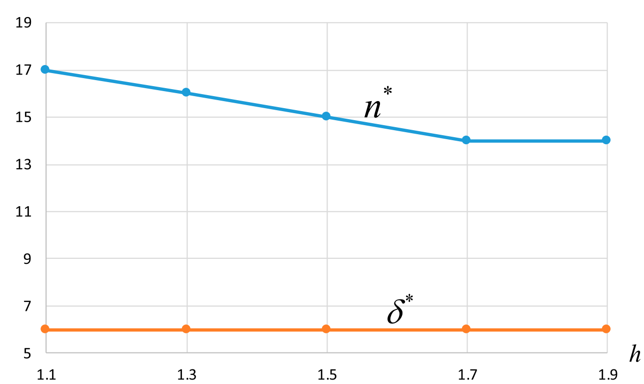

Figure 5. Optimal n and δ as a function of h.

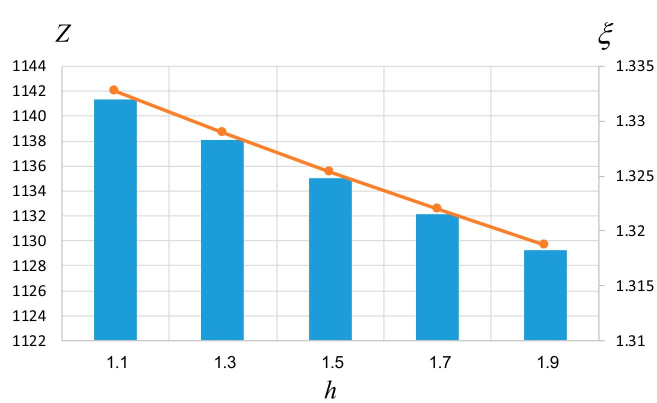

Figure 6. $Z(n^*, \delta^*)$ (bars) and ξ (curve) as a function of h.

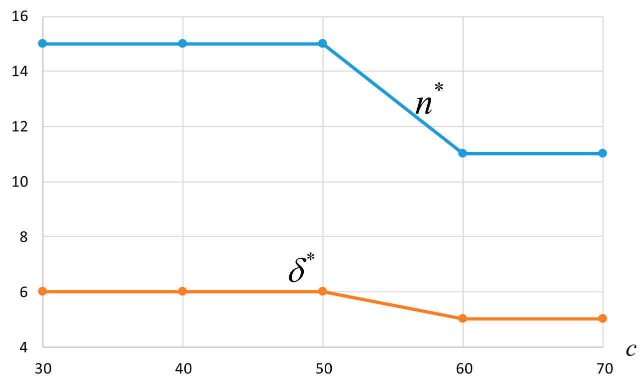

Figure 7. Optimal n and δ as a function of c.

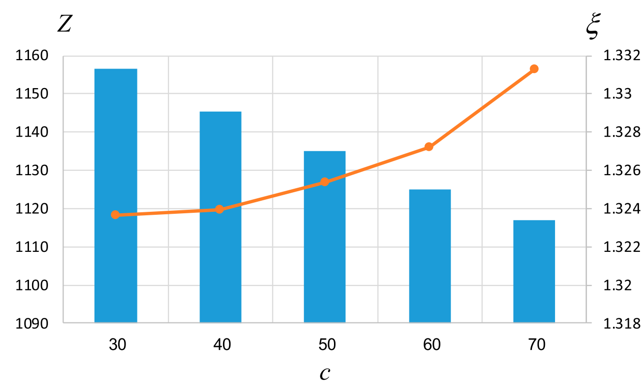

Figure 8. $Z(n^*, \delta^*)$ (bars) and ξ (curve) as a function of c.

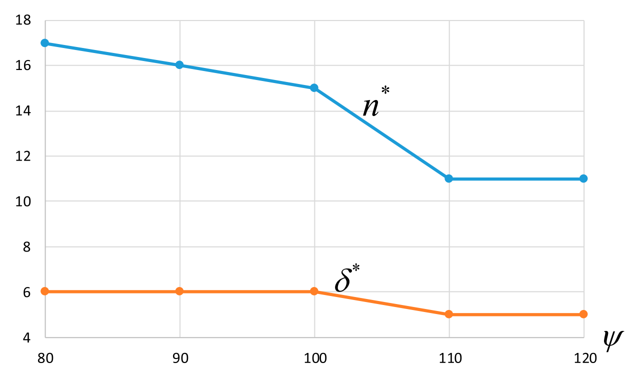

Figure 9. Optimal n and δ as a function of ψ.

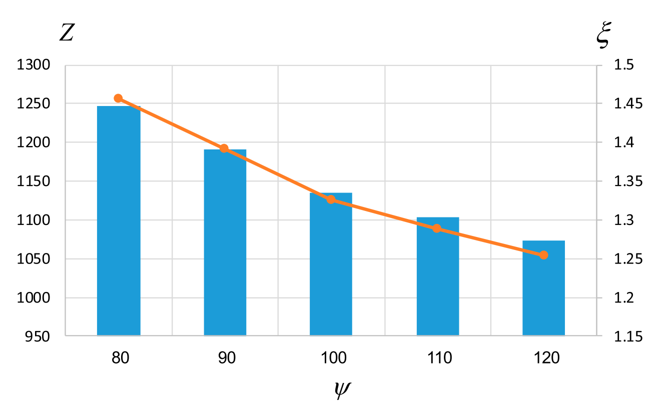

Figure 10. $Z(n^*, \delta^*)$ (bars) and ξ (curve) as a function of ψ.

Figures 3, 5, 7 and 9 show that the optimal value of n (δ) is relatively sensitive (robust) to changes in the model's parameter values; however, it was shown in Table 1 that the reward is robust (sensitive) to the value of n (δ) in the neighbourhood of its optimal value. Therefore, we recommend for the store owner to make an effort to achieve the optimal values of n and δ, and only if the model parameters are subject to considerable changes should a policy modification be considered.

7. Extension to perishable PSs

We now consider the case where the stored preliminary services are subject to decay, which we assume follows a Markovian process; i.e. the lifetime of a PS is exponentially distributed with parameter θ. The transition-rate diagram that includes the deterioration rate is depicted in Figure 11.

In this model, matrix Q takes on the same form as in Equation (1). Matrices A_0 and A_2 remain the same as in the previous model, whereas matrix $A_1 \equiv [a_1^{v,t}]$ is given by

$$a_1^{v,t} = \begin{cases} -(\lambda + 2\mu) & v = t = 1 \\ -(\lambda + \mu + \beta) & v = t = 2 \\ -(\lambda + 2\beta + \theta(v-3)/2) & v = t = 3, 5, 7, \ldots, 2n-1 \\ -(\lambda + \mu + \beta + \theta(v-2)/2) & v = t = 4, 6, 8, \ldots, 2n \\ \theta(v-3)/2 & v = 3, 5, 7, \ldots, 2n-1;\ t = v-2 \\ \theta(v-2)/2 & v = 4, 6, 8, \ldots, 2n;\ t = v-2 \\ 0 & \text{Otherwise} \end{cases}.$$

Matrices $B_{i,j}$, $i \neq j$, remain the same as in the previous model, whereas matrices $B_{i,j}$, $i = j$, are given in Appendix 4. Since matrices A_0, A_1 and A_2 are still all lower triangular, and since, as before, $a_0^{1,1} = \lambda$, $a_2^{1,1} = 2\mu$, the system's stability condition remains $\lambda < 2\mu$, as in the case of the regular M/M/2 queue. The rate matrix R is calculated directly via the formulae given in

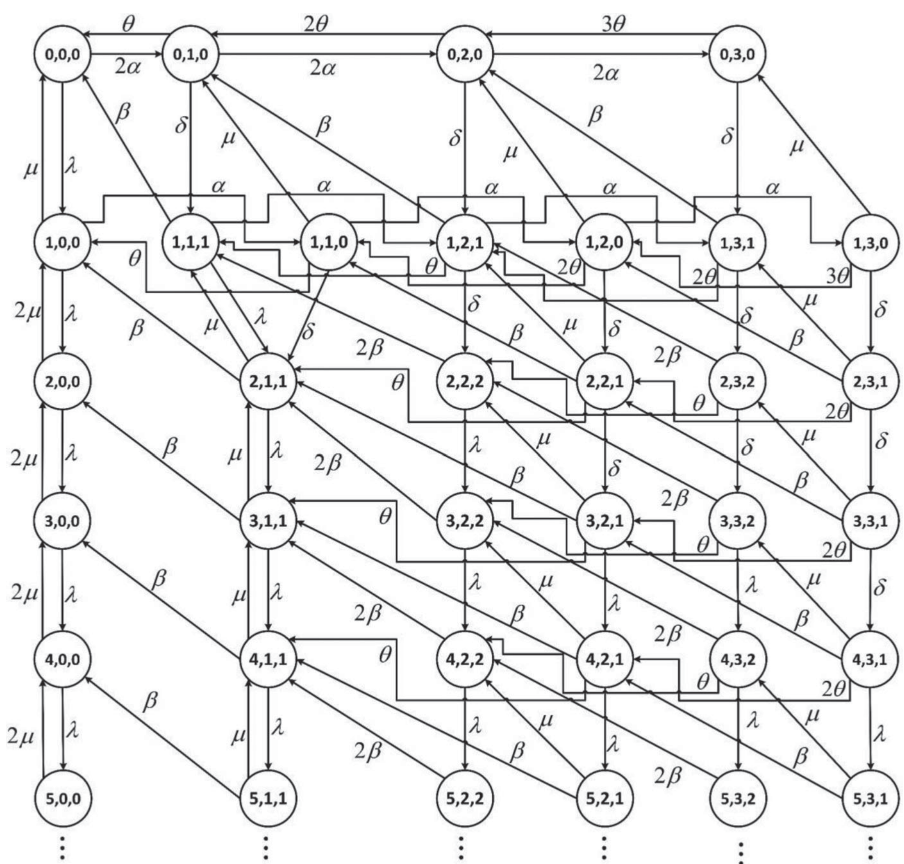

Figure 11. System states and transition-rate diagram for $n = 3$ with deterioration in the preliminary services.

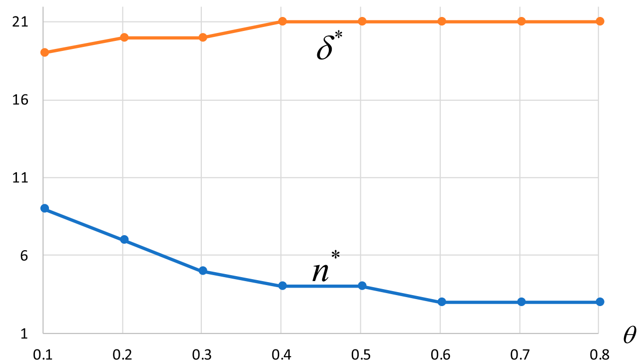

Figure 12. Optimal n and δ as a function of θ.

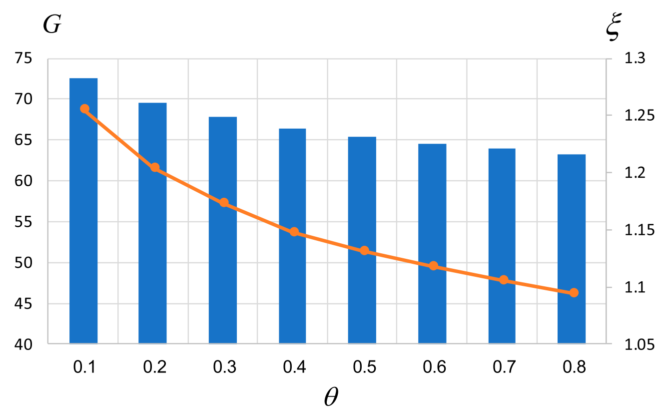

Figure 13. $G(n^*, \delta^*)$ (bars) and ξ (curve) as a function of θ.

Appendix 2. By substituting the matrices $R, A_i, B_{i,j}, \forall i, j$, into Equation (6) and solving the set, the steady-state probabilities are calculated. The performance measures are then calculated by substituting these probabilities into Equations (7)–(13). Let θ_{eff} be the effective deterioration rate, which is calculated by $\theta_{eff} = \theta S_q(n)$, and let k be the loss associated with a spoiled PS. The objective is to maximise the total expected profit per time unit, G, by controlling the PS capacity n and the desired arrival rate δ. Thus, the optimisation problem is defined as follows:

$$\max_{\substack{n\in\{0,1,2\ldots\}\\ \delta\geq\lambda}} \{G(n,\delta) = REV\lambda_{eff}(n,\delta) - A(n,\delta) - cL(n,\delta) - hn - k\theta_{eff}(n,\delta)\}. \tag{15}$$

In order to illustrate the optimisation process and conduct a sensitivity analysis with respect to the added parameter θ, we use a numerical example of a pizza store that uses similar parameter values to those in Hanukov et al. (2020). In particular, we consider a two-server pizzeria in which customers arrive according to a Poisson process with rate $\lambda = 16[\frac{\text{customers}}{\text{hour}}]$, where each customer buys one pizza. It takes on average six minutes ($\mu = 10[\frac{\text{customers}}{\text{hour}}]$) to prepare a pizza according to the customers' requirements. However, when the server prepares a pizza during his idle time, it takes five minutes ($\alpha = 12[\frac{\text{PSs}}{\text{hour}}]$) to prepare the pizza and another one minute to warm it up after a customer buys it ($\beta = 60[\frac{\text{CSs}}{\text{hour}}]$).

We further assume that a pizza is sold for $15[\frac{\$}{\text{customer}}]$,[5] and we estimate the net revenue as one third of the selling price, i.e. $REV = 5[\frac{\$}{\text{customer}}]$. Thus, the cost due to a spoiled pizza is estimated to be $k = 10[\frac{\$}{\text{customer}}]$. We estimate the sojourn cost of a customer to be $c = 5[\frac{\$}{\text{hour}\times\text{customer}}]$ (due to loss of goodwill) and the capacity cost to be $h = 0.25[\frac{\$}{\text{hour}\times\text{PS}}]$ (the alternative cost of using the surface to prepare other products for sale). In addition, we use $\psi = 1\left[\frac{\$\times(\text{hour})^{\tau-1}}{(\text{customer})^{\tau}}\right]$ and $\tau = 1.6$, implying that the promotion required to attract each buyer costs, per hour, at least 20% of the net revenue from selling a pizza.

From Figure 12 we conclude that the optimal investment in promoting the existence of available PSs, which is positively correlated with δ^*, is robust to changes in the value of θ, whereas the optimal PS capacity n^* decreases with θ. The latter relation can be explained by the willingness to avoid loss due to spoilage, which, in this example, is higher than the loss due to the waiting time of customers. Figure 13 shows that, as expected, the maximal expected profit G decreases with θ, and since the expected profit of M/M/2 is independent of θ, then the profit ratio ξ also decreases with θ. We conclude that the reward may change by up to 15% due to the perishability effect, and the store owner may wish to invest in preservation equipment to mitigate this effect (see, for example, Hanukov et al. 2019a).

8. Conclusions

This paper analyses a multi-server system in which customers' arrival and service rates follow Markovian processes. We show that utilising the servers' idle time to produce preliminary services for incoming customers, thus stimulating demand by creating the anticipation of a shorter sojourn time, results in increasing the reward of the service system's owner. Two case studies demonstrate that the system's owner may obtain significant profit gains by using the proposed operational and marketing strategy. In particular, we observe an increase of up to 45.7% in the profit of a bike store and an increase of up to 25.6% in that of a pizza store. The lower increase in the latter case is due to the perishability effect in the food industry, which reduces the benefit from producing preliminary services that may eventually become spoilt and hence would need be scrapped. From a managerial perspective, our models can be applied to many industries in which a service includes a product tailored to the specific requirements of customers. Moreover, our model can serve as an approximation to more complex cases.

For a Markovian model, closed-form expressions for the steady-state probabilities of the system states are derived by applying matrix geometric analysis, which facilitates the straightforward calculation of various performance measures. This result is notable due to the fact that, in typical applications of matrix geometric analysis, closed-form solutions are rarely possible. Furthermore, it is proved analytically that the stability condition of our model is identical to that of a standard M/M/2 queue. This implies that even with the preparation of PSs, the system can hold a load as if it was a regular M/M/2 queue, while the sojourn time of customers is lower and thus the service is considered superior.

We suggest the following directions for future research: First, it would be of great value to relax the Markovian assumptions and study more general arrival, production and service processes, such as PH distributions. Second, the dichotomous effect of the stock of PSs on demand could be extended to a continuous effect. Finally, a worthwhile avenue for further research would be to consider different types of customer, namely regular, fastidious (who consume only fresh products), and strategic (who may balk if the queue length is too high).

Notes

1. https://monopoli.co.il/blog-business/how-to-start-business/%D7%90%D7%99%D7%9A-%D7%9C%D7%A4%D7%AA%D7%95%D7%97-%D7%97%D7%A0%D7%95%D7%AA-%D7%90%D7%95%D7%A4%D7%A0%D7%99%D7%99%D7%9D-%D7%91%D7%99%D7%A9%D7%A8%D7%90%D7%9C.
2. https://bikepanel.com/%d7%94%d7%aa%d7%90%d7%92%d7%93%d7%95%d7%aa-%d7%97%d7%a0%d7%95%d7%99%d7%95%d7%aa-%d7%94%d7%90%d7%95%d7%a4%d7%a0%d7%99%d7%99%d7%9d-%d7%94%d7%a4%d7%a8%d7%98%d7%99%d7%95%d7%aa-%d7%94%d7%90%d7%9d-%d7%96/.
3. https://www.quora.com/How-long-does-it-take-to-assemble-a-bike-from-a-box.
4. https://www.amazon.com/bikes/b?ie = UTF8&node = 1265458011.
5. https://www.forbes.com/sites/priceonomics/2017/04/07/how-much-do-the-ingredients-cost-in-your-favorite-foods/#87c3d9711eda.

Acknowledgements

We thank the two anonymous referees for their constructive comments that helped us to improve our paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the Israel Science Foundation [grant number 1448/17].

ORCID

Gabi Hanukov http://orcid.org/0000-0002-0117-6804

References

Abouee-Mehrizi, H., O. Baron, O. Berman, and D. Chen. 2019. "Managing Perishable Inventory Systems with Multiple Priority Classes." *Production and Operations Management* 28 (9): 2305–2322.

Adacher, L., and C. G. Cassandras. 2014. "Lot Size Optimization in Manufacturing Systems: The Surrogate Method." *International Journal of Production Economics* 155: 418–426.

Altendorfer, K., and S. Minner. 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., and W. H. Zijm. 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. 2020. "The Effect of Decision Rights Allocation on a Supply Chain of Perishable Products Under a Revenue-Sharing Contract." *Accepted for Publication in the International Journal of Production Economics*. doi:10.1016/j.ijpe.2019.107587.

Avinadav, T., and T. Arponen. 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., T. Chernonog, Y. Lahav, and U. Spiegel. 2017. "Dynamic Pricing and Promotion Expenditures in an EOQ Model of Perishable Items." *Annals of Operations Research* 248: 75–91.

Avinadav, T., A. Herbon, and U. Spiegel. 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., A. Herbon, and U. Spiegel. 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.

Baek, J. W., H. W. Lee, S. W. Lee, and S. Ahn. 2014. "A Workload Factorization for BMAP/G/1 Vacation Queues Under Variable Service Speed." *Operations Research Letters* 42: 58–63.

Barron, Y. 2018. "A Threshold Policy in a Markov-Modulated Production System with Server Vacation: The Case of Continuous and Batch Supplies." *Advances in Applied Probability* 50 (4): 1246–1274.

Barron, Y. 2019. "A State-Dependent Perishability (*s, S*) Inventory Model with Random Batch Demands." *Annals of Operations Research* 280 (1–2): 65–98.

Barron, Y., and O. Baron. 2020. "QMCD Approach for Perishability Models: The (*S, s*) Control Policy with Lead Time." *IISE Transactions* 52 (2): 133–150.

Benjaafar, S., W. L. Cooper, and S. Mardan. 2011. "Production-Inventory Systems with Imperfect Advance Demand Information and Updating." *Naval Research Logistics* 58 (2): 88–106.

Berk, E., and Ü Gürler. 2008. "Analysis of the (*Q*, *r*) Inventory Model for Perishables with Positive Lead Times and Lost Sales." *Operations Research* 56 (5): 1238–1246.

Boxma, O. J., S. Schlegel, and U. Yechiali. 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.

Chao, X., X. Gong, C. Shi, and H. Zhang. 2015. "Approximation Algorithms for Perishable Inventory Systems." *Operations Research* 63 (3): 585–601.

Chebolu-Subramanian, V., and G. M. Gaukler. 2015. "Product Contamination in a Multi-Stage Food Supply Chain." *European Journal of Operational Research* 244 (1): 164–175.

Chen, X., Z. Pang, and L. Pan. 2014. "Coordinating Inventory Control and Pricing Strategies for Perishable Products." *Operations Research* 62 (2): 284–300.

Chernonog, T. 2020. "Inventory and Marketing Policy in a Supply Chain of a Perishable Product." *International Journal of Production Economics* 219: 259–274.

Chernonog, T., and T. Avinadav. 2017. "Pricing and Advertising in a Supply Chain of Perishable Products Under Asymmetric Information." *International Journal of Production Economics* 209: 249–264.

Cooper, W. L. 2001. "Pathwise Properties and Performance Bounds for a Perishable Inventory System." *Operations Research* 49 (3): 455–466.

Cox, D. R., and H. D. Miller. 1965. *The Theory of Stochastic Processes*. Wiley.

Dolgui, A., M. K. Tiwari, Y. Sinjana, S. K. Kumar, and Y. J. Son. 2018. "Optimising Integrated Inventory Policy for Perishable Items in a Multi-Stage Supply Chain." *International Journal of Production Research* 56 (1-2): 902–925.

Flapper, S. D. P., J. P. Gayon, and S. Vercraene. 2012. "Control of a Production–Inventory System with Returns Under Imperfect Advance Return Information." *European Journal of Operational Research* 218 (2): 392–400.

Guha, D., V. Goswami, and A. D. Banik. 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., and R. Hassin. 2011. "Strategic Behavior and Social Optimization in Markovian Vacation Queues." *Operations Research* 59 (4): 986–997.

Guo, P., and Z. G. Zhang. 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, S. K. 1967. "Queues with Hyper-Poisson Input and Exponential Service Time Distribution with State Dependent Arrival and Service Rates." *Operations Research* 15 (5): 847–856.

Hanukov, G., T. Avinadav, T. Chernonog, U. Spiegel, and U. Yechiali. 2017. "A Queueing System with Decomposed Service and Inventoried Preliminary Services." *Applied Mathematical Modelling* 47: 276–293.

Hanukov, G., T. Avinadav, T. Chernonog, U. Spiegel, and U. Yechiali. 2018. "Improving Efficiency in Service Systems by Performing and Storing 'Preliminary Services'." *International Journal of Production Economics* 197: 174–185.

Hanukov, G., T. Avinadav, T. Chernonog, and U. Yechiali. 2019a. "Performance Improvement of a Service System via Stocking Perishable Preliminary Services." *European Journal of Operational Research* 274 (3): 1000–1011.

Hanukov, G., T. Avinadav, T. Chernonog, and U. Yechiali. 2019b. "A Multi-Server Queueing-Inventory System with Stock-Dependent Demand." *IFAC-PapersOnLine* 52 (13): 671–676.

Hanukov, G., T. Avinadav, T. Chernonog, and U. Yechiali. 2020. "A Service System with Perishable Products Where Customers are Either Fastidious or Strategic." *International Journal of Production Economics* 228, doi:10.1016/j.ijpe.2020.107696.

Hanukov, G., and U. Yechiali. 2020. "Explicit Solutions for Continuous-Time QBD Processes by Using Relations Between Matrix Geometric Analysis and the Probability Generating Functions Method." *Probability in Engineering and Informational Sciences*, 1–16. doi:10.1017/S0269964819000470.

Herbon, A. 2017. "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. 2018. "Optimal two-Level Piecewise-Constant Price Discrimination for a Storable Perishable Product." *International Journal of Production Research* 56 (5): 1738–1756.

âĂŔHerbon, A., and E. Khmelnitsky. 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.

Hu, P., S. Shum, and M. Yu. 2015. "Joint Inventory and Markdown Management for Perishable Goods with Strategic Consumer Behavior." *Operations Research* 64 (1): 118–134.

Hwang, J., L. Gao, and W. Jang. 2010. "Joint Demand and Capacity Management in a Restaurant System." *European Journal of Operational Research* 207 (1): 465–472.

Ioannidis, S., O. Jouini, A. A. Economopoulos, and V. S. Kouikoglou. 2013. "Control Policies for Single-Stage Production Systems with Perishable Inventory and Customer Impatience." *Annals of Operations Research* 209 (1): 115–138.

Iravani, S. M., B. Kolfal, and M. P. Van Oyen. 2011. "Capability Flexibility: A Decision Support Methodology for Parallel Service and Manufacturing Systems with Flexible Servers." *IIE Transactions* 43 (5): 363–382.

Jeganathan, K., M. A. Reiyas, S. Padmasekaran, and K. Lakshmanan. 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.

Kırcı, M., I. Biçer, and R. W. Seifert. 2019. "Optimal Replenishment Cycle for Perishable Items Facing Demand Uncertainty in a two-Echelon Inventory System." *International Journal of Production Research* 57 (4): 1250–1264.

Koshy, T. 2008. *Catalan Numbers with Applications*. Oxford: Oxford University Press.

Krishnamoorthy, A., R. Manikandan, and B. Lakshmy. 2015. "A Revisit to Queueing-Inventory System with Positive Service Time." *Annals of Operations Research* 233 (1): 221–236.

Lacomme, P., A. Moukrim, A. Quilliot, and M. Vinot. 2018. "Supply Chain Optimisation with Both Production and Transportation Integration: Multiple Vehicles for a Single Perishable Product." *International Journal of Production Research* 56 (12): 4313–4336.

Latouche, G., and V. Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling. ASA-SIAM Series on Statistics and Applied Probability*. Philadelphia, PA: SIAM.

Levy, Y., and U. Yechiali. 1975. "Utilization of Idle Time in an M/G/1 Queueing System." *Management Science* 22: 202–211.

Levy, Y., and U. Yechiali. 1976. "An M/M/s Queue with Servers' Vacations." *INFOR* 14: 153–163.

Li, Q., P. Yu, and X. Wu. 2016. "Managing Perishable Inventories in Retailing: Replenishment, Clearance Sales, and Segregation." *Operations Research* 64 (6): 1270–1284.

Mytalas, G. C., and M. A. Zazanis. 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., M. J. Jacob, and A. Krishnamoorthy. 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*. BaltimoreâĂŔ: Johns Hopkins University Press.

Otten, S., R. Krenzle, and H. Daduna. 2016. "Models for Integrated Production-Inventory Systems: Steady State and Cost Analysis." *International Journal of Production Research* 54 (20): 6174–6191.

Rosenberg, E., and U. Yechiali. 1993. "The M^X/G/1 Queue with Single and Multiple Vacations Under the LIFO Service Regime." *Operations Research Letters* 14 (3): 171–179.

Urban, T. L. 2005. "Inventory Models with Inventory-Level-Dependent Demand: A Comprehensive Review and Unifying Theory." *European Journal of Operational Research* 162 (3): 792–804.

Wang, K., Z. Jiang, N. Li, and N. Geng. 2013. "Optimal Production and Admission Control for a Stochastic SOM System with Demands for Product and PSS." *International Journal of Production Research* 51 (23–24): 7270–7288.

Wang, K., S. Lan, and Z. Jiang. 2016. "Impact of Customer Impatience on a Production Service System." *International Journal of Production Research* 54 (9): 2731–2749.

Wang, J., and X. Zhang. 2017. "Optimal Pricing in a Service-Inventory System with Delay-Sensitive Customers and Lost Sales." *International Journal of Production Research* 55 (22): 6883–6902.

Winston, W. 1978. "Optimality of Monotonic Policies for Multiple-Server Exponential Queuing Systems with State-Dependent Arrival Rates." *Operations Research* 26 (6): 1089–1094.

Yang, D. Y., and C. H. Wu. 2015. "Cost-minimization Analysis of a Working Vacation Queue with N-Policy and Server Breakdowns." *Computers & Industrial Engineering* 82: 151–158.

Yechiali, U. 2004. "On the M^X/G/1 Queue with a Waiting Server and Vacations." *Sankhya* 66: 159–174.

Zhang, H., C. Shi, and X. Chao. 2016. "Approximation Algorithms for Perishable Inventory Systems with Setup Costs." *Operations Research* 64 (2): 432–440.

Zhao, N., and Z. Lian. 2011. "A Queueing-Inventory System with two Classes of Customers." *International Journal of Production Economics* 129 (1): 225–231.

Appendices

Appendix 1

$$B_{1,1}=\begin{pmatrix}-(\lambda+2\alpha) & 2\alpha & 0 & \cdots & 0\\ 0 & -(\delta+2\alpha) & 2\alpha & & 0\\ 0 & 0 & -(\delta+2\alpha) & \ddots & \vdots\\ \vdots & \vdots & & \ddots & 2\alpha\\ 0 & 0 & 0 & \cdots & -\delta\end{pmatrix}_{(n+1)\times(n+1)}$$

$$B_{2,3}=\begin{pmatrix}\lambda & 0 & 0 & 0 & \cdots & 0\\ 0 & \lambda & 0 & 0 & \cdots & 0\\ 0 & \delta & 0 & 0 & \cdots & 0\\ 0 & 0 & \delta & 0 & & 0\\ 0 & 0 & 0 & \delta & & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & 0 & \cdots & \delta\end{pmatrix}_{(2n+1)\times 2n} \qquad B_{1,2}=\begin{pmatrix}\lambda & 0 & 0 & 0 & \cdots & 0 & 0\\ 0 & \delta & 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & 0 & \delta & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & 0 & 0 & \cdots & \delta & 0\end{pmatrix}_{(n+1)\times(2n+1)}$$

$$B_{2,1}=\begin{pmatrix}\mu & 0 & 0 & \cdots & 0 & 0\\ \beta & 0 & 0 & \cdots & 0 & 0\\ 0 & \mu & 0 & \cdots & 0 & 0\\ 0 & \beta & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & \mu & 0\\ 0 & 0 & 0 & \cdots & \beta & 0\\ 0 & 0 & 0 & \cdots & 0 & \mu\end{pmatrix}_{(2n+1)\times(n+1)}$$

$$B_{2,2}=\begin{pmatrix}-(\lambda+\mu+\alpha) & 0 & \alpha & 0 & 0 & \cdots & 0 & 0\\ 0 & -(\lambda+\beta+\alpha) & 0 & \alpha & 0 & \cdots & 0 & 0\\ 0 & 0 & -(\delta+\mu+\alpha) & 0 & \alpha & \cdots & 0 & 0\\ 0 & 0 & 0 & -(\delta+\beta+\alpha) & 0 & \cdots & 0 & 0\\ 0 & 0 & 0 & 0 & -(\delta+\mu+\alpha) & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & 0 & 0 & 0 & \cdots & \alpha & 0\\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \alpha\\ 0 & 0 & 0 & 0 & 0 & \cdots & -(\delta+\beta) & 0\\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & -(\delta+\mu)\end{pmatrix}$$

$$B_{3,1}=\begin{pmatrix}2\mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0\\ \beta & \mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0\\ 0 & 2\beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & \beta & \mu & 0 & 0 & 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & 2\beta & 0 & 0 & 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & 0 & \beta & \mu & 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & 0 & 0 & 2\beta & 0 & 0 & 0 & & \vdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & & & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \beta & \mu & 0\end{pmatrix}_{2n\times(2n+1)}$$

The square matrices $B_{i,2} \equiv [b_{i,2}^{v,t}]$, $i = 3, 4, 5, \ldots, n+1$, of size $2n \times 2n$, are given by

$$b_{i,2}^{v,t} = \begin{cases} -(\lambda + 2\mu) & v = t = 1 \\ -(\lambda + \mu + \beta) & v = t = 2, 4, 6, 8, \ldots, 2i-4 \\ -(\lambda + 2\beta) & v = t = 3, 5, 7, \ldots, 2i-3 \\ -(\delta + \mu + \beta) & v = t = 2i-2, 2i, 2i+2, \ldots, 2n \\ -(\delta + 2\beta) & v = t = 2i-1, 2i+1, 2i+3, \ldots, 2n-1 \end{cases}.$$

Similarly,

$$B_{i,3} \equiv [b_{i,3}^{v,t}],$$

where

$$b_{i,3}^{v,t} = \begin{cases} \lambda & v = t = 1, 2, 3 \ldots, 2i-3 \\ \delta & v = t = 2i-2, 2i-1, 2i, \ldots, 2n \end{cases}.$$

Appendix 2

$$r^{v,t} = 0 \quad \text{for } v < t, \tag{A2.1}$$

i.e. R is lower triangular,

$$r^{v,v} = \begin{cases} \dfrac{-a_1^{v,v} - \sqrt{(a_1^{v,v})^2 - 4\, a_0^{v,v}\, a_2^{v,v}}}{2\, a_2^{v,v}}, & a_2^{v,v} > 0,\ a_0^{v,v} > 0 \\ 0, & a_2^{v,v} > 0,\ a_0^{v,v} = 0 \\ \dfrac{-a_0^{v,v}}{a_1^{v,v}}, & a_2^{v,v} = 0 \end{cases}, \quad \forall v. \tag{A2.2}$$

$$r^{v,t} = -\frac{a_0^{v,t} + \sum_{k=t+1}^{v} r^{v,k} a_1^{k,t} + \sum_{\tau=t+1}^{v-1} r^{v,\tau} r^{\tau,t} a_2^{t,t} + \sum_{k=t+1}^{v} \sum_{\tau=k}^{v} r^{v,\tau} r^{\tau,k} a_2^{k,t}}{a_1^{t,t} + a_2^{t,t}(r^{t,t} + r^{v,v})} \quad \text{for } v > t. \tag{A2.3}$$

Appendix 3

Since $a_0^{v,t} = 0$ for $v \neq t$, $a_1^{k,t} = 0$ for $k \neq t$ and $a_2^{t,t} = 0$ for $t > 2$, then Equation (A2.3) reduces to

$$r^{v,t} = -\frac{\sum_{k=t+1}^{v} \sum_{\tau=k}^{v} r^{v,\tau} r^{\tau,k} a_2^{k,t}}{a_1^{t,t}}. \tag{A3.1}$$

(i) Proof of the first term of (vi)

When $t = 4, 6, 8, \ldots, 2n$, then $a_1^{t,t} = -(\beta + \mu + \lambda)$, $a_2^{t+2,t} = \beta$, and $a_2^{k,t} = 0$ for $k \neq t+2$. Substituting these values into (A3.1) leads to

$$r^{v,t} = \frac{\beta \sum_{\tau=t+2}^{v} r^{v,\tau} r^{\tau,t+2}}{\lambda + \mu + \beta}. \tag{A3.2}$$

We now prove by induction that $r^{v,t} = 0$ for $t = 4, 6, 8, \ldots, 2n-2$, $v = j+1, j+2, \ldots, 2n-1$. First, $r^{v,t} = 0$ for $t = 2n-2$, $v = 2n-1$, which is proved by substituting the values of t and v into (A3.2), which eliminates the summation. We now assume that $r^{v,t} = 0$ for $t \geq 2n-2k$, $v = t+1, t+3, \ldots, 2n-1$ and show that $r^{v,t} = 0$ for $t = 2n-2(k+1)$, $v = t+1, t+3, \ldots, 2n-1$. By substituting the value of t into the summation in (A3.2) we get $\sum_{\tau=2n-2k}^{v} r^{v,\tau} r^{\tau,2n-2k}$. By the induction assumption, when τ is even, $r^{v,\tau} = 0$, and when τ is not even, $r^{\tau,2n-2k} = 0$, which proves the claim.

(ii) Proof of the second term of (vi)

We now prove by induction that

$$r^{v,t} = \frac{C_{0.5(v-t)} \beta^{0.5(v-t)} \lambda^{0.5(v-t+2)}}{(\lambda + \mu + \beta)^{v-t+1}} \quad \text{for } t = 4, 6, 8, \ldots, 2n, v = t, t+2, t+4, \ldots, 2n. \tag{A3.3}$$

First, (A3.3) is proved for $v = t = 2n$ by substituting $a_0^{2n,2n} = \lambda$ and $a_1^{2n,2n} = -(\lambda + \mu + \beta)$ into the third term of (A2.2), which is used since $a_2^{2n,2n} = 0$. We now assume (A3.3) for $t \geq 2n-2k$, $v = t, t+2, t+4, \ldots, 2n$ and show that it is valid for $t = 2n-2(k+1)$, $v = t, t+2, t+4, \ldots, 2n$. By substituting the latter value of t into (A3.2), and by considering that when τ is not even, $r^{\tau,2n-2k} = 0$, we get

$$r^{v,2n-2(k+1)} = \frac{\beta \sum_{\substack{\tau=2n-2k \\ \tau\ even}}^{v} r^{v,\tau} r^{\tau,2n-2k}}{\beta+\mu+\lambda}$$
$$= \frac{\beta^{0.5(v-2n+2(k+1))}\lambda^{0.5(v-2n+2(k+1)+2)} \sum_{\substack{\tau=0 \\ \tau\ even}}^{v-(2n-2k)} C_{0.5(v-(2n-2k)-\tau)} C_{0.5\tau}}{(\lambda+\mu+\beta)^{v-(2n-2k)+1}}$$

By rearranging the indices in the summation and by using the recurrence relation of Catalan numbers, $\sum_{k=0}^{m} C_k C_{m-k} = C_{m+1}$, we get

$$\sum_{\substack{\tau=0 \\ \tau\ even}}^{v-(2n-2k)} C_{0.5(v-(2n-2k)-\tau)} C_{0.5\tau} = \sum_{\tau=0}^{0.5(v-(2n-2k))} C_{0.5(v-(2n-2k))-\tau} C_\tau = C_{0.5(v-(2n-2(k+1)))},$$

which proves the claim.

(iii) Proof of the second term of (vii)

When $t = 3, 5, 7, \ldots, 2n-1$, then $a_1^{t,t} = -(\lambda+2\beta)$, $a_2^{t+2,t} = 2\beta$, $a_2^{t+1,t} = \mu$, and $a_2^{k,t} = 0$ for $k \neq t+1, t+2$. Substituting these values into (A3.1) leads to

$$r^{v,t} = \frac{\mu \sum_{\tau=t+1}^{v} r^{v,\tau} r^{\tau,t+1} + 2\beta \sum_{\tau=t+2}^{v} r^{v,\tau} r^{\tau,t+2}}{2\beta+\lambda}. \tag{A3.4}$$

We now prove by induction that

$$r^{v,t} = \frac{2^{0.5(v-t)} C_{0.5(v-t)} \beta^{0.5(v-t)} \lambda^{0.5(v-t+2)}}{(\lambda+2\beta)^{v-t+1}} \quad \text{for } t = 3, 5, 7, \ldots, 2n-1, v = t, t+2, \ldots, 2n-1. \tag{A3.5}$$

First, (A3.5) is proved for $v = t = 2n-1$ by substituting $a_0^{2n-1,2n-1} = \lambda$ and $a_1^{2n-1,2n-1} = -(\lambda+2\beta)$ into the third term of (A2.2), which is used since $a_2^{2n-1,2n-1} = 0$. We now assume (A3.5) for $t \geq 2n-1-2k$, $v = t, t+2, t+4, \ldots, 2n-1$ and show that it holds for $t = 2n-1-2(k+1)$, $v = t, t+2, t+4, \ldots, 2n-1$. By substituting the latter value of t into (A3.4), and by considering that $r^{v,\tau} = 0$ when τ is even and $r^{\tau,t+1} = 0$ when τ is not even, we get

$$r^{v,2n-1-2(k+1)} = \frac{2\beta \sum_{\substack{\tau=2n-1-2k \\ \tau\ not\ even}}^{v} r^{v,\tau} r^{\tau,2n-1-2k}}{\lambda+2\beta}$$
$$= \frac{2^{0.5(v-(2n-1-2(k+1)))} \beta^{0.5(v-(2n-1-2(k+1)))} \lambda^{0.5(v-(2n-1-2(k+1))+2)} \sum_{\substack{\tau=2n-1-2k \\ \tau\ not\ even}}^{v} C_{0.5(v-\tau)} C_{0.5(\tau-(2n-1-2k))}}{(\lambda+2\beta)^{v-(2n-1-2(k+1))+1}}$$

By rearranging the indices in the summation, we get

$$\sum_{\substack{\tau=2n-1-2k \\ \tau\ not\ even}}^{v} C_{0.5(v-\tau)} C_{0.5(\tau-(2n-1-2k))} = \sum_{\tau=0}^{0.5(v-(2n-1-2k))} C_{0.5(v-(2n-1-2k))-\tau} C_\tau = C_{0.5(v-(2n-1-2(k+1)))},$$

which proves the claim.

Appendix 4

$$B_{1,1} = \begin{pmatrix} -(\lambda+2\alpha) & 2\alpha & 0 & \cdots & 0 & 0 \\ \theta & -(\delta+2\alpha+\theta) & 2\alpha & \cdots & 0 & 0 \\ 0 & 2\theta & -(\delta+2\alpha+2\theta) & \cdots & 0 & 0 \\ 0 & 0 & 3\theta & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\delta+2\alpha+(n-1)\theta) & 2\alpha \\ 0 & 0 & 0 & \cdots & n\theta & -(\delta+n\theta) \end{pmatrix}_{(n+1)\times(n+1)}$$

$$B_{2,2} = \begin{pmatrix}
-(\lambda+\mu+\alpha) & 0 & \alpha & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \\
0 & -(\lambda+\beta+\alpha) & 0 & \alpha & 0 & \cdots & 0 & 0 & 0 & 0 & \\
\theta & 0 & -(\delta+\mu+\alpha+\theta) & 0 & \alpha & \cdots & 0 & 0 & 0 & 0 & \\
0 & \theta & 0 & -(\delta+\beta+\alpha+\theta) & 0 & \cdots & 0 & 0 & 0 & 0 & \\
0 & 0 & 2\theta & 0 & -(\delta+\mu+\alpha+2\theta) & \cdots & 0 & 0 & 0 & 0 & \\
\vdots & \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & -(\delta+\beta+\alpha+(n-2)\theta) & 0 & \alpha & 0 & \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 & -(\delta+\mu+\alpha+(n-1)\theta) & 0 & \alpha & \\
0 & 0 & 0 & 0 & 0 & \cdots & (n-1)\theta & 0 & -(\delta+\beta+(n-1)\theta) & 0 & \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 & n\theta & 0 & -(\delta+\mu+n\theta) &
\end{pmatrix}$$

The matrix $B_{i,2} \equiv [b_{i,2}^{v,t}]$ is given by

$$b_{i,2}^{v,t} = \begin{cases}
-(\lambda+2\mu) & v=t=1 \\
-(\lambda+\mu+\beta) & v=t=2 \\
-(\lambda+2\beta+\theta(v-3)/2) & v=t=3,5,7,\dots,2i-3 \\
-(\delta+2\beta+\theta(v-3)/2) & v=t=2i-1,2i+1,2i+3,\dots,2n-1 \\
-(\lambda+\mu+\beta+\theta(v-2)/2) & v=t=4,6,8,\dots,2i-4 \\
-(\delta+\mu+\beta+\theta(v-2)/2) & v=t=2i-2,2i,2i+2,\dots,2n \\
\theta(v-3)/2 & v=3,5,7,\dots,2n-1;\ t=v-2 \\
\theta(v-2)/2 & v=4,6,8,\dots,2n;\ t=v-2 \\
0 & \text{Otherwise}
\end{cases}.$$