Matrix-Analytic Methods in Supply Chain Management: Recent Developments

Robert N. Boute
Pieter J. Colen
Stefan Creemers
Ann Noblesse
Benny Van Houdt

Abstract - **Matrix-analytic methods are a popular modeling tool in a great number of fields, most notable in the analysis of telecommunication systems. Because of their ability to construct and analyze a wide class of stochastic models, they can also be applied in the analysis of complex supply chain problems where traditional analytical techniques or simulation analysis fall short. In this paper, we demonstrate the power of matrix-analytic methods in the analysis of four different supply chain problems: (1) to determine lead times in production/inventory models characterized by any arbitrary discrete (i.e., non-Poisson) demand distribution; (2) to gain insight in the upstream replenishment orders driven by (s, S) inventory policies; (3) to analyze waiting times and resource utilization in service systems that are driven by appointments (e.g., health care, legal services, administration); and (4) to determine the optimal maintenance policy/warranty in the aftermarket supply chain.**

Keywords - **Markov chain analysis, supply chain management, stochastic modeling**

1 Introduction: matrix-analytic methods

Over the last three decades, broad classes of frequently encountered queueing models have been analyzed by *matrix-analytic methods* [15, 16, 12]. The Markov chains in these models are two-dimensional generalizations of the classic M/G/1 and GI/M/1 queues, and birth-death processes. Consider a discrete-time Markov chain (MC) with transition matrix

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & \dots \\ C_{-1} & A_0 & A_1 & A_2 & \dots \\ C_{-2} & A_{-1} & A_0 & A_1 & \ddots \\ C_{-3} & A_{-2} & A_{-1} & A_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where A_i, for all i, B_i, for $i \geq 0$, and C_{-i}, for $i \geq 1$, are nonnegative matrices in $\mathbb{R}^{m \times m}$ such that $C_{-j} + \sum_{i=-j+1}^{+\infty} A_i$, for $j \geq 1$, and $\sum_{i=0}^{+\infty} B_i$ are stochastic. Such an MC is termed

a GI/M/1-type MC if $A_i = B_i = 0$, for $i \geq 2$, and an M/G/1-type MC if $A_{-i} = C_{-i} = 0$, for $i \geq 2$. MCs belonging to the class of GI/M/1-type and M/G/1-type MCs are called Quasi-Birth-Death (QBD) MCs, their only non-zero blocks are $A_{-1}, A_0, A_1, B_0, B_1$ and C_{-1}. The main computational problem is to compute the invariant probability vector of P (if it exists), i.e., the infinite nonnegative row vector π such that $\pi P = \pi$ and $\pi e = 1$, where e is the vector with all its entries equal to 1. Denote $\pi = (\pi_0, \pi_1, \ldots)$ such that π_i is a $1 \times m$ vector, for $i \geq 0$.

For GI/M/1-type MCs, [15] has proven that the vector π has a matrix-geometric form, that is, $\pi_{i+1} = \pi_i R$, for $i \geq 1$, where $R \in \mathbb{R}^{m \times m}$ is the minimal nonnegative solution to the nonlinear matrix equation

$$R = \sum_{i=-1}^{+\infty} R^{i+1} A_{-i}. \tag{1}$$

Note that if the MC belongs to the class of QBDs, Eq. (1) reduces to $R = A_1 + RA_0 + R^2 A_{-1}$.

In the M/G/1-type case, π does not have a matrix-geometric form, but can be expressed via Ramaswami's formula [see 12] which expresses π_{i+1} in terms of π_0 to π_i and in terms of a matrix $G \in \mathbb{R}^{m \times m}$ which is the minimal nonnegative solution to the nonlinear matrix equation

$$G = \sum_{i=-1}^{+\infty} A_i G^{i+1}. \tag{2}$$

Various algorithms have been introduced to compute matrices R and G iteratively, such as functional iterations, cyclic reduction, the invariant subspace approach, the Newton iteration and logarithmic reduction (for QBDs only). We refer to [3, 2, 4] who have implemented these algorithms in an SMCSolver software tool.

Apart from the GI/M/1-type, M/G/1-type and QBD MCs, matrix-analytic models include notions such as the Markovian arrival process (MAP) and the phase-type (PH) distribution, both in discrete and continuous time. A discrete-time PH distribution X corresponds to the time until absorption in an $n+1$-state discrete-time MC with a transition matrix of the form

$$P = \begin{bmatrix} T & t \\ 0 & 1 \end{bmatrix},$$

with T substochastic and given that the initial state is generated according to the probability vector $(\alpha, 0)$. As the rows of P must be stochastic, $t = e - Te$. Hence, a discrete PH distribution is characterized by the triple (n, T, α) and $\Pr[X = k] = \alpha T^{k-1} t$, for $k \geq 1$. Continuous-time PH distributions are defined in a similar manner by means of a continuous-time MC with a single absorbing state such that $\Pr[X \leq t] = 1 - \alpha \exp(Tt)t$, for $t \geq 0$. The class of PH distributions contains many well known distributions as special cases, such as hyper-exponential/geometric, Erlang and Coxian distributions, as well as any distribution with finite support on $\{0, 1, 2, \ldots\}$. Further, the class of continuous-time distributions is also dense in the class of all distributions on $[0, \infty)$, meaning that any distribution can be approximated arbitrarily close by means of a PH distribution. Various tools for fitting PH distributions have been developed (e.g., BuTools, PHfit and ProFiDo).

Matrix-analytic methods have been applied in various areas such as telecommunication systems, production, computer engineering, inventory modeling, ruin theory, health care,

just to name a few. In this paper we demonstrate its strength on four different supply chain management problems.

2 Lead times in general discrete production and/or inventory systems

Much of the management science literature separates the questions of production and inventory control. Most inventory models in the literature take the replenishment lead time as a fixed constant or as an exogenous variable with a given probability distribution. Treating production and inventory as independent units may be reasonable, when e.g., the inventory and production systems belong to separate entities, and the production entity guarantees fixed delivery times, or when, e.g., transportation times are significantly longer than manufacturing lead times. However, in many firms or integrated supply chains, inventory directly influences production by initiating orders and loading the production facilities, and production in turn influences inventory by completing and delivering orders to inventory. Therefore, the inventory control system should work with a lead time which is a good estimate of the effective supply lead time. [22, p. 246] states: *"to understand the overall inventory system, we need to understand the supply system. For this purpose we can and do apply the results of queueing theory"*.

To cope with this, we consider an integrated production/inventory (P/I) model, where the inventory control system generates replenishment orders that are sent to production; the time to replenish the order is determined by the sojourn time or response time in the queue and in production. In this model, supply lead times depend on the production load, the arrival rate of orders and the variability of the production system. These lead times are a prime determinant in setting safety stock requirements.

2.1 Modeling approach

We assume a single item P/I system with random period demands, and inventory levels reviewed periodically and managed using a base-stock policy (this is a common supply chain setting). Under these assumptions, the arrival process of orders at the queue is characterized by batch arrivals with a fixed interarrival time, equal to the review period, and with variable batch sizes, equal to the deficit between base-stock level and inventory position. When the review period is one base period and demand is IID, the replenishment orders are discrete IID random variables with the same general distribution as the demand pattern. As a result we obtain a $D^G/G/1$ queue, which are much more complex to analyse than the (frequently adopted) $M/M/1$ queues.

To analyse the lead times in a $D^G/G/1$ queue, we make use of matrix-analytic methods and PH distributions. Remember that any general distribution can be approximated in sufficient detail by means of a PH distribution. The key idea behind PH distributions is to exploit the Markovian structure of the distribution to simplify the queueing analysis. When we fit both the demand pattern and the single unit service times by a discrete PH distribution, we obtain a $D^{PH}/PH/1$ queue. Moreover, when both the batch size and the single unit service process are PH distributed, then the production time of an entire batch (an

integral replenishment order) is also PH distributed characterized by the triple (n_S, T_S, α_S). This permits us to treat the entire batch order as a single service unit and hence, the problem of estimating the lead time is reduced to computing the response time distribution of a unit in a $D/PH/1$ queue. The $D/PH/1$ queue is a special case of the $PH/PH/1$ queue and can be analysed numerically using matrix-analytic methods as follows.

To compute the response time T_r of a batch order in a $D/PH/1$ queue, we construct a MC (B_n, S_n), where B_n represents the age of the order in service at the n-th observation point t_n and S_n reflects the phase of the service process at epoch t_n. The age of the order in service at time t_n is defined as the duration of the time interval $[a_n, t_n)$, where a_n denotes the arrival time of the replenishment order (i.e., the time the order was placed). Instead of observing the MC (B_n, S_n) at all time epochs, we observe the system only when the server is busy (simplifying the boundary behavior of the MC). Hence, the time of the n-th observation point, t_n, is the n-th epoch during which the server is busy.

The MC (B_n, S_n) has an infinite number of states labeled $1, 2, \ldots$. The set of states $\{(i-1)n_S + 1, \ldots, in_S\}$ is referred to as level i of the MC, for $i \geq 1$. The states of level $i > 0$ are labeled as s, where $1 \leq s \leq n_S$. Let state s of level i of the MC correspond to the situation in which there is an order in service (being produced), that arrived i time units ago, while the service process is currently in phase s. The transition matrix P of this MC can be written as a GI/M/1-type MC:

$$P = \begin{bmatrix} A_d & A_0 & 0 & \ldots & 0 & 0 & \ldots \\ A_d & 0 & A_0 & \ldots & 0 & 0 & \ldots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots \\ A_d & 0 & 0 & \ldots & A_0 & 0 & \ldots \\ 0 & A_d & 0 & \ldots & 0 & A_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \ddots \end{bmatrix}.$$

The matrix A_0 represents the probabilities that the service of the batch continues and is given by $A_0 = T_S$: when an order remains in service, the level (age) of the batch goes up with one time unit, and the new state (service phase) is given by the PH distributed service process T_S. The matrix A_d represents the probabilities that the service of the batch finishes and is given by $A_d = t_S \alpha_S$ (with $t_S = e - T_S e$): when a batch order is completed (with probability t_S), the next order starts service, which arrived one review period ago (so its age is one review period less). The phase of the service process is now given by α_S. If an order completes service before the next order is placed, then a transition is made to level 1, since our MC is defined only when the server is busy.

The steady state vector π of P can be found using the rate matrix R (see Eq. (1)). Having obtained the steady state vector $\pi = (\pi_1, \pi_2, \ldots)$, we can obtain the response time T_r of a batch order using the following observation: The probability that a batch order has a response time of i time units can be calculated as the expected number of orders with an age of i time units that complete their service at an arbitrary time instant, divided by the expected number of orders that complete their service during an arbitrary time instant (that is, $1/d$ for a queue with $\rho < 1$). As such, we find the response time distribution as

$$\Pr[T_r = i] = d\rho\pi_i t_S. \tag{3}$$

2.2 Findings

The procedure using matrix-analytic methods is of importance because in enables to compute the replenishment lead time distribution (and consequently, the distribution of inventory levels, fill rates, base-stock levels and optimal safety stocks) for any arbitrary discrete demand distribution. [5] demonstrate that this rigorous analysis outperforms well-established existing models of capacitated systems: the relaxation of the full distribution assumption (or its two moment approximation) and the relaxation of the endogenous lead time assumption leads to incorrect decisions. Ignoring the exact demand and production characteristics in favor of the more readily available queueing results based on the exponential (geometric) distribution results in a wrong estimation of the safety stock requirements and poor service levels. It shows that a more accurate performance analysis using PH distributions and matrix-analytic methods pays off for the more demanding mathematical analysis.

3 Analysis of upstream order fluctuations in (s, S) models

A frequently used inventory control policy in the inventory management literature is the (s, S) policy. Under this policy, an order is placed to increase the inventory position up to a level S, as soon as it hits the reorder point s. These policies are commonly used in practice when there is a fixed order cost; e.g., because of a fixed setup in production, transportation or administration to process an order; in that case, it makes sense to order in batches with minimum size $Q = S - s$. In the literature, (s, S) inventory policies have been numerically analyzed using MC analysis; e.g., [19] and [20] determine average inventory levels, stock-out probabilities and long-run average costs for a given numerical problem using MCs, and [18] determines the limiting distribution of inventory positions in continuous (s, S) models. In this section, we use the MC approach to characterize the order pattern in a continuous review (s, S) inventory policy. Since this order pattern is transmitted to its upstream supply chain partner, it has a major impact on its upstream supply performance. Therefore, a closer look at the order fluctuations caused by an (s, S) inventory policy is valuable.

In the supply chain literature, it is generally accepted that (s, S) policies and batch ordering lead to increased variability in the supply chain [6]. [13] even denotes batch ordering as one of the root causes of the bullwhip effect. Figure 1 (solid line) illustrates a simulated compound Poisson demand pattern during 150 periods with on average 0.5 customer arrivals per period and Poisson demand sizes with $\mu_{dk} = 8$. Assume a continuous review (s, S) policy with $S - s = 15$ (because of a supposed set-up cost): every time a demand occurs, the inventory position is evaluated. If the inventory position is smaller than or equal to s, an order quantity equal to the difference between S and the inventory position is placed. If it is larger than the order point s, we do not order. One could interpret "not ordering" as "placing an order with an order quantity equal to zero", which leads to an order pattern with highly fluctuating order quantities (see the dotted line in figure 1). However, this is not how the upstream supplier observes this order pattern: the production facility receiving these orders, does not know when a customer demand has occurred and will just observe the time between subsequent orders and the related order quantities. In other words, from

the perspective of the production facility, "not ordering" increases the time between orders and does not count as an order with order quantity equal to zero. Therefore, figure 2 rather than figure 1 reflects how the upstream supplier looks at the order pattern. As such, the conjecture of variance amplification (bullwhip), as is always predicted in the literature, does not seem that straightforward anymore (see figure 2).

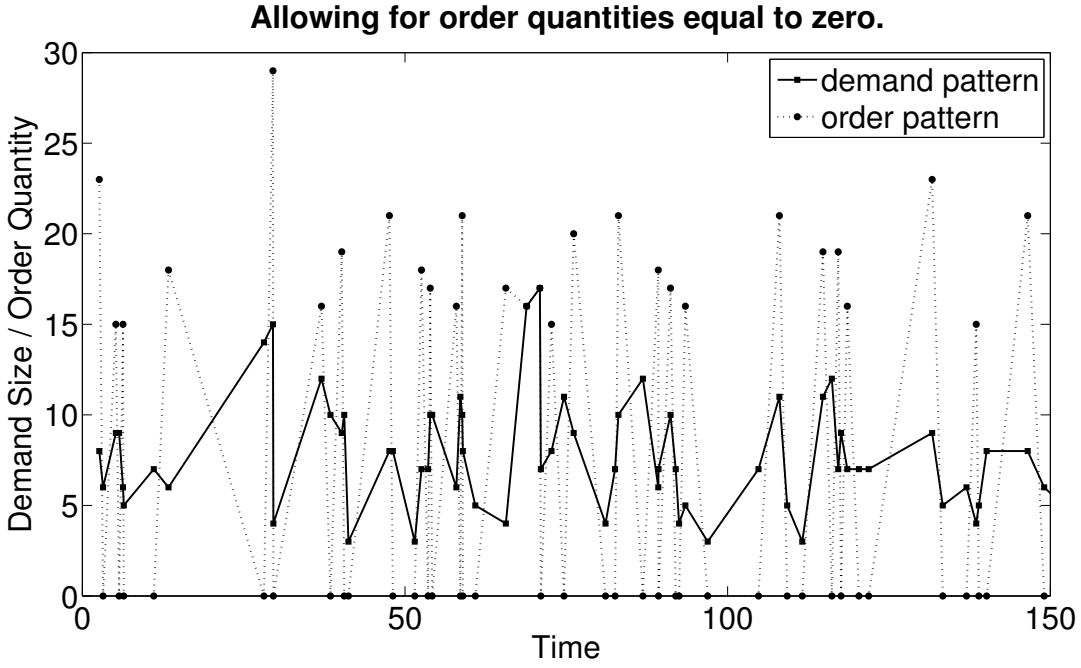


Figure 1: Compound Poisson demand pattern caused by an (s, S) policy

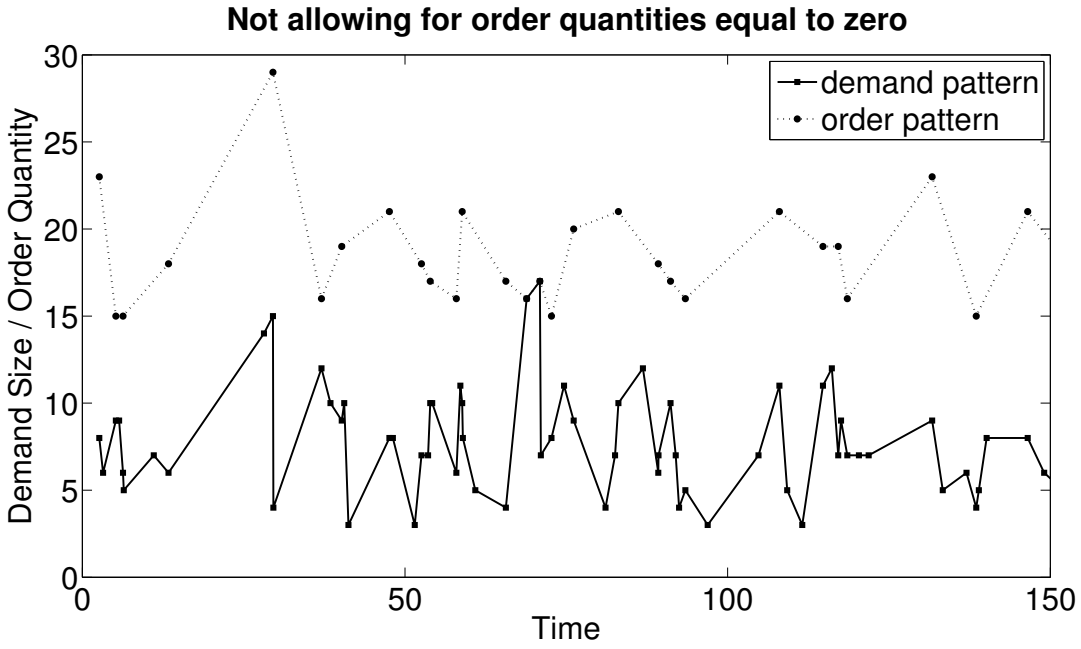


Figure 2: Order pattern caused by an (s, S) policy

Using MCs and the properties of PH distributions, we analyze the order fluctuations caused by (s, S) policies, thereby making a distinction between the fluctuations in order sizes and the fluctuations in time between two orders placed. Throughout this section, we will use the following notation:

λ	rate of demand arrivals
σ_{dt}^2	variance of time between demands
$X(i)$	probability that the demand size is i
σ_{dk}^2	variance of demand sizes
k	order quantity variable
σ_{ok}^2	variance of order quantities
t^*	expected time between orders
σ_{ot}^2	variance of time between orders
Q	minimum order quantity, i.e., $S - s$

3.1 Characterization of order quantities

We use an absorbing continuous-time MC to represent the inventory position in an (s, S) controlled inventory model. The states of the MC refer to the inventory positions prior to replenishment, ranging from S to $-\infty$. Every cycle starts at a replenishment (state S) and a transition to another state occurs when a demand depletes the inventory (in which case the state decreases). The rate at which transitions occur, equals the demand rate λ and the depletion in inventory is given by the demand probabilities $X(i)$. When state s or a lower state is reached, a replenishment is made, and a new cycle starts from state S. Hence we consider states smaller than $s + 1$ as an absorbing state.

The time to reach absorption is equivalent to the time it takes to place the subsequent order. The distribution of this time is equivalent to a continuous-time PH distribution with $S-s+1$ phases, defined by the initial vector α and the transition matrix T. Since every cycle starts at inventory position S, the initial vector α is $\begin{bmatrix}1 & 0 & \dots & 0\end{bmatrix}$. The matrix T describes how the inventory position decreases prior to absorption, with states ranging from S to $s+1$:

$$T = \begin{bmatrix} \lambda \cdot (X(0)-1) & \lambda \cdot X(1) & \lambda \cdot X(2) & \dots & \lambda \cdot X(Q-1) \\ 0 & \lambda \cdot (X(0)-1) & \lambda \cdot X(1) & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \lambda \cdot X(2) \\ 0 & 0 & 0 & \ddots & \lambda \cdot X(1) \\ 0 & 0 & 0 & \ddots & \lambda \cdot (X(0)-1) \end{bmatrix}.$$

We can use the properties of PH distributions to analyze the time spent in each inventory position during one cycle (i.e., until replenishment occurs). We refer to $\frac{1}{\lambda}b_i$ as the time spent in inventory position i, given that $i > s$. This time is given by $\alpha(-T)^{-1}$. After normalizing, i.e., dividing by the total time spent in all inventory positions prior to absorption, we derive the steady state probabilities v_j of inventory position j, for $j \geq s+1$:

$$v_j = b_j / \left[\sum_{i=0}^{Q-1} b_{S-i}\right] \tag{4}$$

$$b_S = \frac{1}{1-X(0)} \tag{5}$$

$$b_{S-j} = \frac{1}{1-X(0)} \sum_{i=1}^{j} X(i) \cdot b_{S-j+i}, \tag{6}$$

with $Q = S-s$. Eq. (4) allows us to find $\pi(k)$, the steady-state probability to order quantity k. For instance, the probability to order quantity Q is equal to the probability that the inventory position exactly reaches the order point s. This probability is given by the sum for all i (ranging from 0 to $Q-1$) of being in inventory position $S-i$ and a demand of size $Q-i$ realizes, in which case the inventory position drops to s and a replenishment of size Q is placed. Similarly, the probability to order quantity k, with $k = Q, Q+1, \dots, \infty$, can be defined as:

$$\pi(k) = \left[\sum_{i=0}^{Q-1} X(k-i) \cdot b_{S-i}\right] / \left[\sum_{k=Q}^{\infty} \sum_{i=0}^{Q-1} X(k-i) \cdot b_{S-i}\right]. \tag{7}$$

The variance in order quantities is then defined by:

$$\sigma_{ok}^2 = \left[\sum_{k=Q}^{\infty} (k)^2 \cdot \pi(k)\right] - \left[\sum_{k=Q}^{\infty} (k) \cdot \pi(k)\right]^2. \tag{8}$$

The difference between the order quantity k and Q is called the *overshoot*. In case of only single-unit demands, there is no overshoot, and Eqs. (7-8) reduce to $\pi(Q) = 1$ and $\sigma_{ok}^2 = 0$.

If $Q = 1$ (and demand can be greater than one), Eqs. (7-8) reduce to: $\pi(k) = X(k)$ and $\sigma_{ok}^2 = \sigma_{dk}^2$. We obtain an order-up-to policy with the orders equal to demand.

3.2 Characterization of time between orders placed

In the previous section we illustrated how we can determine the fluctuations in order sizes. Either variance amplification or dampening is possible, depending on the demand size distribution and the minimum order quantity Q. In this section, we provide some general results to characterize the time between orders placed.

Proposition 1 *The average time between two subsequent orders t^* is larger than or equal to the average time between two demand arrivals $1/\lambda$:*

$$t^* \geq \frac{1}{\lambda}, \text{ with } t^* = \frac{1}{\lambda}\sum_{j=0}^{Q-1} b_{S-j}. \tag{9}$$

Proof: The average time to place the subsequent order t^* is determined by the sum of the average time spent in all inventory positions prior to the order point s, which are given by $\frac{1}{\lambda}b_i$. As all b_i are positive and $b_S \geq 1$, we find that $t^* \geq \frac{1}{\lambda}$. ■

Proposition 2 *The variance of the time between subsequent orders is larger than or equal to the variance of time between subsequent arrivals of demand:*

$$\sigma_{ot}^2 \geq \sigma_{dt}^2, \text{ with } \sigma_{ot}^2 = \frac{1}{\lambda^2}\left[\left(2\sum_{j=0}^{Q-1} b_{S-j}\sum_{i=0}^{Q-1-j} b_{S-i}\right) - \left(\sum_{j=0}^{Q-1} b_{S-j}\right)^2\right]. \tag{10}$$

Proof: In case of single-unit demands, the probabilities to be in inventory positions $S, S-1, \ldots, s+1$ are all equal to $1/Q$ [see also 18]. The time between two orders is the sum of Q exponential interarrival times between subsequent demands and its distribution follows an Erlang distribution. It follows that $\sigma_{ot}^2 = \frac{Q}{\lambda^2}$, and for $Q \geq 1$, $\sigma_{ot}^2 \geq \sigma_{dt}^2$.

For general demands, the time between two orders is the sum of a variable number of exponential interarrival times between demands, and therefore PH distributed. It has been shown that an Erlang distribution is the least variable type of PH distributions [1]. This implies that, in general, $\sigma_{ot}^2 \geq \frac{Q}{\lambda^2}$. As $Q \geq 1$ (the case $Q = 1$ refers to an order-up-to policy), it follows that $\sigma_{ot}^2 \geq \sigma_{dt}^2$. ■

Proposition 3 *The squared coefficient of variation of time between subsequent orders is smaller than or equal to the squared coefficient of variation of time between subsequent arrivals of demand:*

$$\frac{\sigma_{ot}^2}{(t^*)^2} \leq \frac{\sigma_{dt}^2}{\left(\frac{1}{\lambda}\right)^2} \tag{11}$$

Proof: As demand arrives according to a compound Poisson process, it follows that the squared coefficient of variation (scv) of the time between subsequent demand arrivals equals 1. Hence, to prove Proposition 3, we need to prove that $\frac{\sigma_{ot}^2}{(t^*)^2} \leq 1$, or $\sigma_{ot}^2 \leq (t^*)^2$.

Substituting Eq. (9) and Eq. (10), we need to show that

$$\frac{1}{\lambda^2}\left[\left(2\sum_{j=0}^{Q-1} b_{S-j}\sum_{i=0}^{Q-1-j} b_{S-i}\right) - \left(\sum_{j=0}^{Q-1} b_{S-j}\right)^2\right] \leq \left[\frac{1}{\lambda}\sum_{j=0}^{Q-1} b_{S-j}\right]^2, \tag{12}$$

or by rewriting Eq. (12),

$$\sum_{j=0}^{Q-1} b_{S-j} \left(\sum_{i=0}^{Q-1-j} b_{S-i} - \sum_{i=0}^{Q-1} b_{S-i} \right) \leq 0. \tag{13}$$

As all b_i are positive, we know that $\sum_{i=0}^{Q-1-j} b_{S-i} \leq \sum_{j=0}^{Q-1} b_{S-j}$, which proves Proposition 3. ■

3.3 Findings

Using the properties of PH distributions in the MC analysis of continuous review (s, S) inventory policies, we are able to characterize the time between orders and the order quantities. Let us go back to figure 2: if demand sizes are Poisson distributed with $\mu_{dk} = 8$, the arrival rate of customers $\lambda = 0.5$, and $S - s = 15$, we find that: (1) the squared coefficient of variation (scv) of order quantities is 0.0254, whereas the scv of demand sizes is 0.125; and (2) the scv of time between orders equals 0.4768, whereas the scv of time between demands equals 1. If scv is our measure to evaluate variability, we find that there is no variance amplification or bullwhip in the order quantities and in time between orders in this particular example.

We characterized the fluctuations of the order pattern in a realistic way from the point of view of the supplier (i.e., not taking zero order quantities into account). These fluctuations of the order pattern impact the upstream supply performance. In further research we will analyse their impact on supply lead times (using the method of [21]), and look at the order pattern in a compound way, i.e., combining the fluctuations in order sizes and interarrival times in a *compound* measure. At that point in time, we will be able to state whether the order pattern (combination of order quantities and time between orders) is more variable compared to the demand pattern (combination of demand sizes and time between demands).

4 Appointment-driven queueing systems

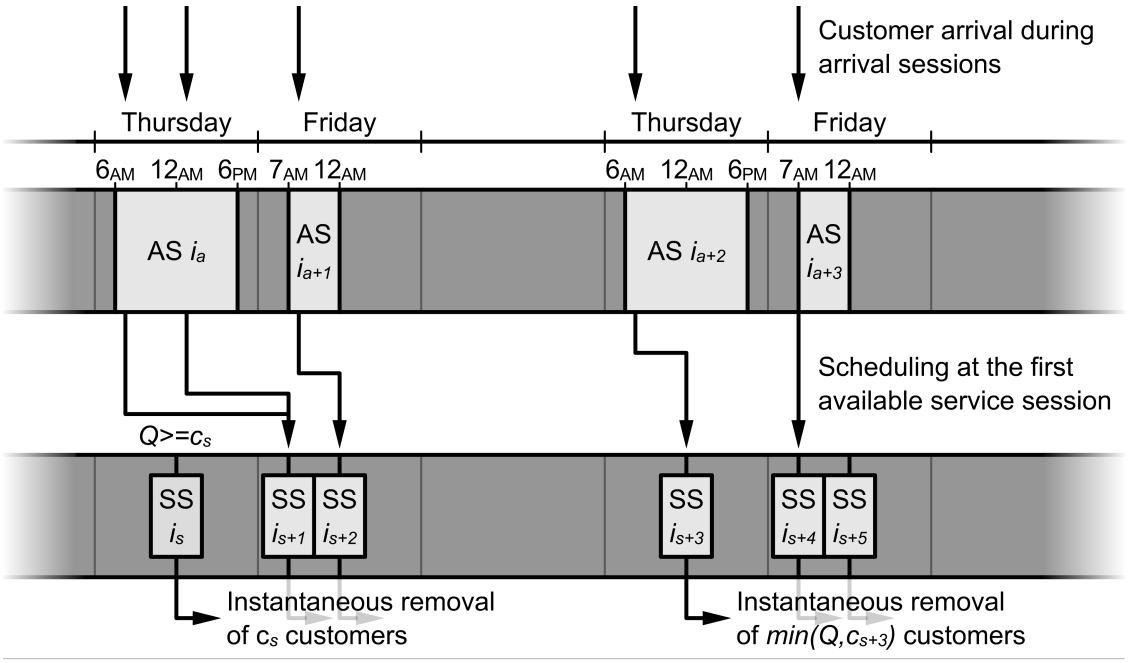
Many service systems require customers to make an appointment prior to receiving service. Often, these appointments are made during so-called "arrival sessions" (e.g., during office hours). After making an appointment, customers are issued an appointment date (which takes place at some future "service session") and join an external queue which is referred to as the "waiting list". Upon the appointment date, the customer is removed from the waiting list and receives service at the "service facility" (e.g., a doctors office). We refer to these systems as appointment-driven systems. They may be found in health care, legal services, administration and many other service and manufacturing industries.

We are interested in the queueing behavior of a customer from the moment an appointment is made (i.e., the arrival of a customer) until the moment that service is received. We do not focus on what happens during the service session itself (we refer to [9] for models in which these elements are incorporated). To do so, we analyse a bulk service queueing model that assigns customers to the first upcoming service session that has capacity available. The properties of the bulk service queueing model may be summarized as follows:

- Customers arrive during arrival sessions.

- Customers are removed from the queue only at the start of a service session.
- Customers receive service in the first upcoming service session in which capacity is available.
- Customers are removed instantaneously, in batches and according to a FCFS policy.
- The number of customers removed depends on the current queue size and on the maximum number of customers that is allowed to receive service during the upcoming service session.

Figure 3 illustrates the dynamics of this bulk service queueing system. In this particular example we assume that service takes place during "service sessions"on Thursday (at 12 AM) and on Friday (at 7 AM and at 12 AM). Customers arrive during "arrival sessions"on Thursday (from 6 AM until 6 PM) and on Friday (from 7 AM until 12 AM). In the example, five customers arrive and are scheduled for service during the first available service session. Observe that the capacity of the first service session is already fully occupied (i.e., the number of customers in the waiting list at the start of the first service session is at least equal to the number of customers that is allowed to receive service during that service session). Therefore, the first arriving customer is scheduled for service at the second service session (in which capacity is still available). At the start of a service session, either all customers are removed from the queue or the maximum number of customers allowed to receive service during the upcoming service session is removed.



AS : Arrival Session
SS : Service Session
Q : Number of customers in queue
c_s : Maximum number of customers served during service session i_s

Figure 3: Illustration of the dynamics of the bulk service queueing system

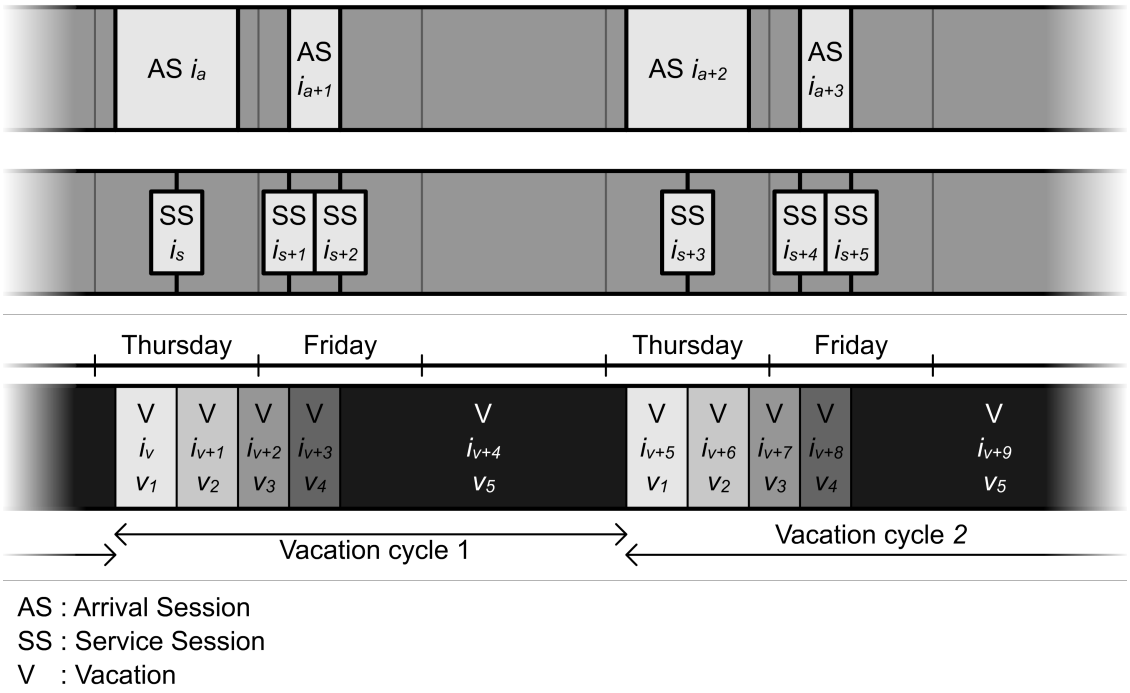


Figure 4: Illustration of the cyclic vacation system

The server operates under a vacation policy. The dynamics of the vacation mechanism are as follows: upon the start of a service session, the server returns from vacation, instantaneously removes up to a maximum number of customers from the queue and departs on a vacation once more. Bulk service queues and vacation models have received a lot of attention during the past decades. For an overview of the literature in the field, refer to [8, 7] and the references therein.

The performance measure of interest is the expected waiting time of a customer in this setting. In order to obtain exact results, we rely on PH distributions and matrix-analytic methods.

4.1 Modeling approach

The service process is a succession of service sessions during which customers are served. Each service session $i_s : s \in \{1, 2, \ldots\}$ is characterized by the maximum number of customers c_s allowed to receive service. We assume recurring cycles to be present in the succession of service sessions (e.g., the use of a specific operating room is assigned to the orthopaedics department every Thursday and Friday afternoon). A cycle of service sessions has length L_s. Similarly to the service process, the arrival process is a succession of arrival sessions $i_a : a \in \{1, 2, \ldots\}$ during which customers are allowed to arrive. We assume recurring cycles to be present in the succession of arrival sessions. A cycle of arrival sessions has length L_a. In building the bulk service queueing model, we will fully exploit the repetitive structure of the service and arrival processes.

The vacation process is obtained when superimposing both the service and the arrival

process. The vacation process is the continuous (i.e., uninterrupted) succession of vacations $i_v : v \in \{1, 2, \ldots\}$ of deterministic length L_v. A new vacation is initiated at each instance in time at which (1) a service session starts; (2) an arrival session starts; and (3) an arrival session ends. This observation is used to determine L_v. Because service and arrival processes are assumed to be cyclic, the vacation process is cyclic as well. The cycle length of the vacation process L_v equals the least common multiple of L_s and L_a (assuming the ratio of L_s and L_a is a rational number). A cycle of vacations contains V vacations. We illustrate these principles in figure 4.

Note that, due to the cyclic nature of the basic processes, a vacation of type $(v + (iV))$ is also a vacation of type v. In addition, vacations may be divided into different classes (e.g., arrivals are allowed to take place only during vacations of a particular class).

The bulk service queueing model presented here is not a straightforward queueing model. One possible approach would be to construct an MC of four dimensions: (1) the queue size $Q : Q \in \{0, 1, 2, \ldots\}$; (2) the vacation type v; (3) the phase of the arrival process $y : y \in \{1, \ldots, Y_v\}$; and (4) the phase of the vacation process $z : z \in \{1, \ldots, Z\}$.

Unfortunately, the use of multidimensional MCs is in general not advisable since it is clear that, as V, Y_v or Z increase, the resulting statespace grows rapidly. When modeling real-life problems, memory- and computational boundaries are easily surpassed. In order to efficiently assess performance measures, we decompose the systems into two subsystems:

- A first subsystem observes the queueing process of customers only at the start of a vacation of type v, prior to the removal of up to c_v customers. We use a set of discrete-time MCs X ($X = \{X_1, \ldots, X_V\}$ and $X_v = \{X_v(t) : t \geq 0\}$) to analyze this first subsystem; where MC X_v may be defined as a two-dimensional stochastic process whose statespace can be represented by pairs $(Q, y)_v$. X_v observes the queueing behavior of customers only at the start of a vacation of type v. The actions taking place in between two successive observation moments are left unobserved (refer to figure 5 for an illustration of this process). From the analysis of MC X_j, we obtain $\mathcal{Q}_v^{(1)}$; the expected number of customers in queue during a vacation of type v, given that these customers did already arrive prior to vacation v.

- A second subsystem observes the queueing process of those customers who arrive during a vacation of type v. Using matrix-analytic methods, we obtain $\mathcal{Q}_v^{(2)}$; the expected number of customers in queue during a vacation of type v, given that these customers did arrive during vacation v.

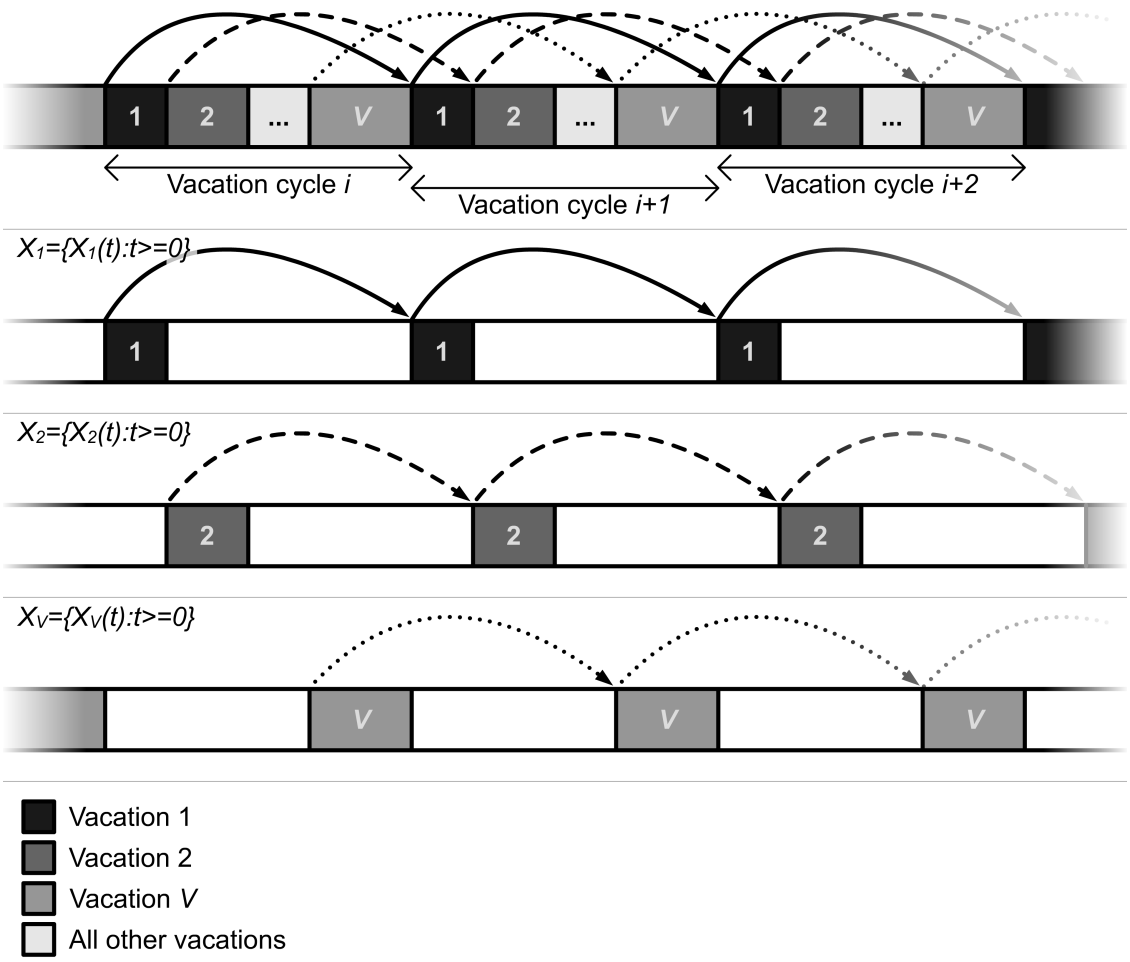


Figure 5: Illustration of vacation jumps

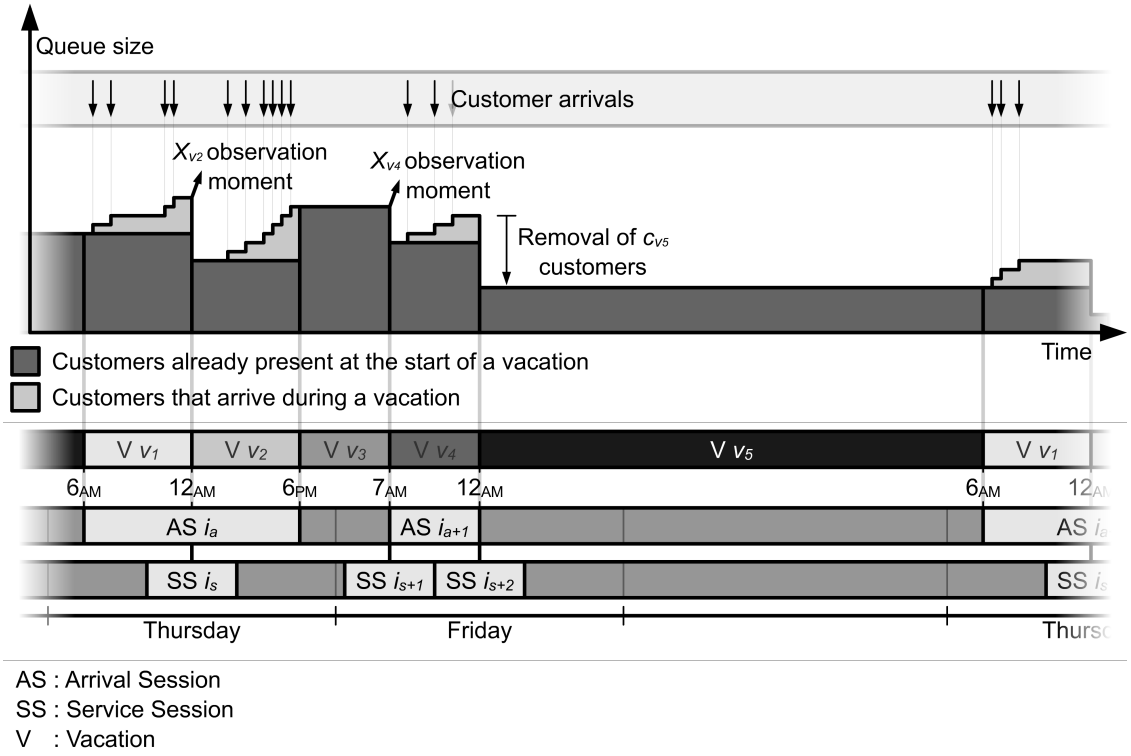


Figure 6: Decomposition of the bulk service queueing system

We illustrate this decomposition in figure 6. Decomposing the system significantly improves computational efficiency due to: (1) dimensional reduction of the statespace; and (2) avoiding unnecessary computations. When aggregating $\mathcal{Q}_v^{(1)}$ and $\mathcal{Q}_v^{(2)}$ over all vacations, we obtain the expected waiting time of a customer.

4.2 Findings

In this section, we used matrix-analytic methods to analyze the queueing behavior of a customer in an appointment-driven system. The presented model allows to optimize the performance of such systems. The model has already been applied to identify robust and efficient appointment scheduling rules and to determine the optimal allocation of server capacity over different classes of patients in a hospital [7]. It is clear, however, that many other applications exist.

5 Maintenance optimization in aftermarket supply chains

As machines are becoming increasingly more complex and technically skilled people are scarce and expensive, many companies have started to outsource the maintenance of their machinery. In line with this market trend, machine manufacturers are now offering a wide range of extended warranty contracts to their customers. Such contracts protect the customer against repair costs but can also include some preventive maintenance (PM). PM is conducted

to avoid failures and requires both parts and technician hours. If a machine fails, parts and technicians need to be deployed to repair the machine. Hence, balancing PM and repair actions (costs) is key for success in the aftermarket. Markovian models can be used to characterize the reliability of machines taking into account the effect of maintenance on the reliability [11, 14].

Our goal is to optimize the maintenance policy of a machine manufacturing company that wants to minimize its warranty costs for a machine type during the selling period of that machine type. When machines are used, their reliability deteriorates (i.e., if the number of operating hours increases, the failure rate increases as well). Machines that fail have to be repaired by the OEM, resulting in additional costs. Because executing repairs is often expensive, it can be beneficial for the OEM to conduct some PM during the warranty period in order to limit the number of failures by slowing down the reliability deterioration process. Different types of maintenance interventions can be executed ranging from light preventive maintenance (A maintenance job), medium preventive maintenance (B job), to very thorough maintenance (C maintenance job). The types of maintenance jobs differ in their impact on machine reliability and incurred costs.

Given both cost and maintenance effectiveness parameters for the different maintenance types, we are able to calculate the impact (costs and reliability) of a specific maintenance policy. A maintenance policy prescribes the timing and type of PM interventions. Hence, we determine the timing and type of jobs to execute in order to minimize the warranty costs, taking into account the effect of maintenance on the expected number of failures. We assume that the maintenance policy is the same for all machines under consideration. Besides providing an efficient and effective method to optimize the maintenance policy during the warranty period, our results make it possible to determine the expected demand for technicians to execute both the preventive and corrective maintenance jobs. A lot of work has been done on maintenance optimization during warranty (e.g., [14, 17, 10]). However, our Markovian modeling approach allows us to model more realistic settings: take into account different types of (imperfect) maintenance interventions (including different repair levels) and deal with large scale problems.

5.1 Modeling Approach

The reliability of machines is reflected by the hazard rate, denoted by $h(t)$, which gives the probability of failure for a machine that has been operating for t hours. We assume that the hazard rate is an increasing function of the operating hours (see figure 7(a)).

In order to model the reliability based on a discrete Markovian process, we split up the continuous failure rate distribution in several failure classes characterized by the average failure rate in the corresponding time interval (see figure 7(a)). At any given moment in time, there are a number of machines that are assigned to failure class 1 (i.e., the newest machines), whereas the older, less reliable machines are assigned to failure class 2 or 3. Besides the hazard rate, we also discretize time: PM can be conducted at discrete, equidistant moments (we denote Δ for the duration between two subsequent moments (e.g., 24 hours)). Clearly, the more failure classes and the shorter the time periods (i.e., Δ), the more accurate the model but the greater the computational effort.

The next step in our model formulation consists of modeling the deterioration process of

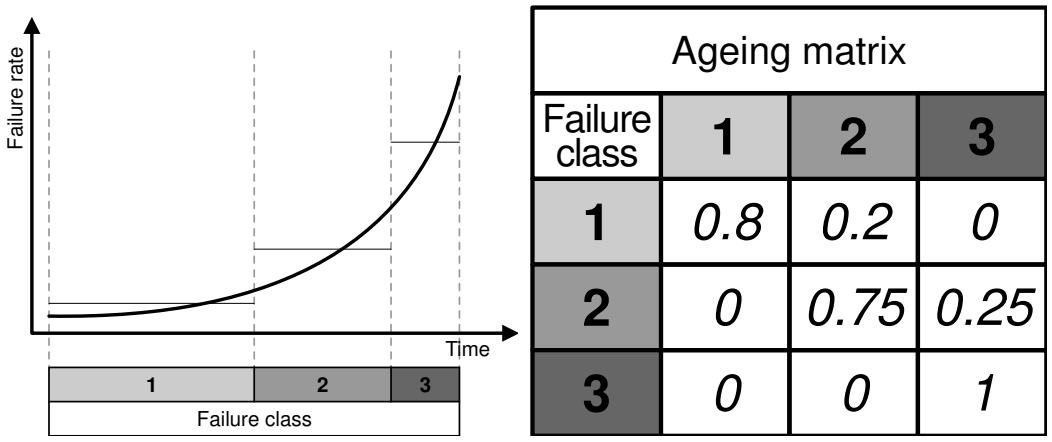


Figure 7: (a) Discretization of the hazard (failure) rate into failure classes; and (b) Modeling the machine aging process

the machines and the impact of PM on the deterioration process. During each time period, the reliability of machines deteriorates which results in a reclassification of some machines into a higher failure class. These deterioration transitions can be represented by means of an "ageing" matrix (refer to figure 7(b) for an example of an ageing matrix).

Figure 7(b) shows, for example, that a machine which is in failure category 1 has a 20% probability to evolve into category 2 after one time period Δ. By conducting PM, the deterioration process can be slowed down. Depending on the type of PM (i.e., job A, B or C), the impact on the deterioration process will be different as will be the costs incurred. Hence, the evolution of the machine park can be represented as shown in figure 8. The variables a, b and c are the fractions of the machine population that are in failure class 1, 2, and 3 respectively at moment t. The second matrix (i.e., the ageing matrix) corresponds with the deterioration process and the impact of PM is incorporated by selecting the appropriate preventive maintenance matrix. The multiplication of the matrices results in the distribution of the machines across the failure classes at moment $t + 1$. If no PM is executed during a certain time period, the preventive maintenance matrix is omitted.

The evolution of the reliability of the machine park can be modeled based on the matrix multiplications as shown in figure 8. Moreover, we can estimate the number of failures during a time period (Δ) for each failure class:

$$E[\text{failures}|\text{machine} \in \text{class } i] = h_i * \Delta, \tag{14}$$

where h_i characterizes the failure rate of class i, and Δ refers to the length of the time period $(t, t + 1)$. Given the costs of conducting the different PM types and the cost of repairs, the total warranty cost for a given maintenance policy can be calculated. This enables the optimization of the maintenance policy by means of a clever algorithm. As such, we identify the optimal maintenance policy that is used to maintain the machine park. Moreover, by taking into account the technician labor hours (or the part requirements) of the different

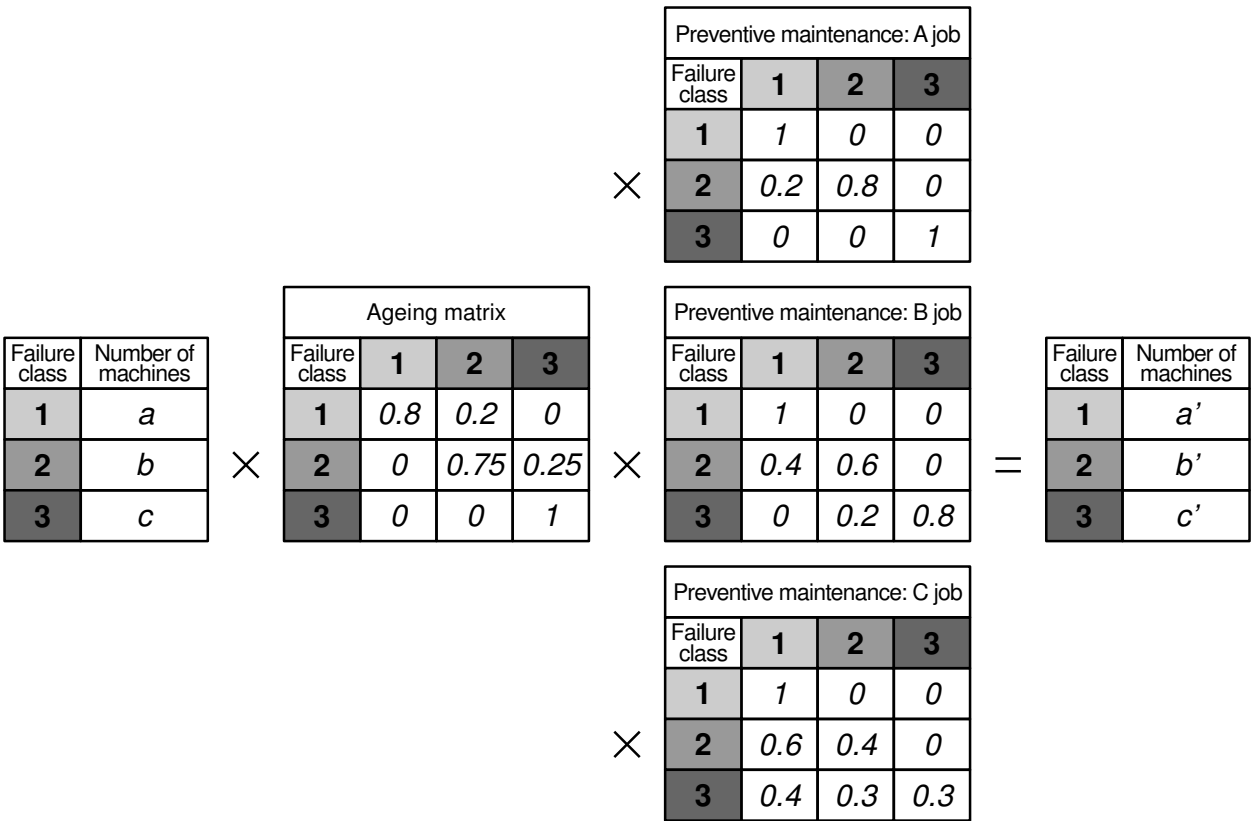


Figure 8: Modeling the evolution of the machine park

maintenance types and repairs, our findings estimate the demand for these critical resources. Knowledge about the expected demand for technicians and parts and its evolution during the time that the machines are being produced, allows more accurate planning of the aftermarket supply chain.

5.2 Findings

We propose a fast and accurate way of optimizing the maintenance policy of a machine park under warranty. Thanks to the Markovian modeling approach, we are able to cope with real-life scaled problems. Our results reveal that in some instances it can be cost-optimal for the equipment manufacturer to bare (some of) the costs of PM during the warranty period. Hence, there is an economic business case to offer customers incentives to conduct PM even during the warranty period by e.g., offering price reductions. Our maintenance policy only considers a limited amount of possible maintenance types instead of a continuous maintenance impact parameter (e.g., [10]). However, in practise companies work with a limited set of possible maintenance interventions and our approach can cope with a large amount of (predefined) maintenance types. [10] also study maintenance practices during the warranty period but from the equipment buyer's perspective and find that also from the buyer's perspective conducting PM during warranty can be optimal. Lastly, we note that our modeling approach can be extended by e.g., incorporating different failure types or

machine sensoring information (i.e., condition based maintenance).

6 Concluding remarks

Modeling techniques typically face a tradeoff between accuracy of the model and the memory- and/or computational requirements. Simulation, for instance, can be used to model virtually any system in a very accurate way; but the price that has to paid comes in the form of long simulation run times. Analytical techniques on the other hand, are often less demanding in terms of computational requirements; unfortunately, most analytical techniques impose restrictive assumptions making them hardly applicable in a more practical setting. In this article, we have shown that matrix-analytic methods are able to combine the best of both worlds: they allow us to analyze complex, real-life systems in an accurate and sometimes even exact manner; at the same time matrix-analytic methods are far less demanding than alternative techniques such as simulation. We demonstrated its strength in four different supply chain settings: (1) to analyse lead times in integrated production/inventory models; (2) to gain insight in the upstream order fluctuations generated by (s, S) inventory policies (3); to analyse waiting times in appointment driven queueing models; and (4) to optimize maintenance policies in aftermarket supply chains. We are convinced that many other supply chain settings exist, in which matrix-analytic methods can prove its value.

References

[1] D. Aldous and L. Shepp. The least variable phase type distribution is erlang. *Stochastic Models*, 3(3):467–473, 1987.

[2] D. Bini, B. Meini, S. Steffé, and B. Van Houdt. Structured markov chain solver: software tools. In *Proc. of the SMCTools workshop*, Pisa, Italy, 2006.

[3] D. A. Bini, G. Latouche, and B. Meini. *Numerical methods for structured Markov chains*. Oxford University Press, 2005.

[4] D. A. Bini, B. Meini, S. Steffé, and B. Van Houdt. Structured Markov chains solver: algorithms. In *SMCtools Workshop*, Pisa, Italy, 2006. ACM Press.

[5] R. N. Boute, M. R. Lambrecht, and B. Van Houdt. Performance evaluation of a production/inventory system with periodic review and endogenous lead times. *Naval Research Logistics*, 54(4):462–473, 2007.

[6] AS Caplin. The Variability of Aggregate Demand with (s,S) Inventory Policies. *Econometrica*, 53(6):1395–1409, 1985.

[7] S. Creemers, J. Beliën, and M.R. Lambrecht. The optimal allocation of server time slots over different classes of patients. *European Journal of Operational Research*, (to appear), 2012.

[8] S. Creemers and M.R. Lambrecht. An advanced queueing model to analyze appointment-driven service systems. *Computers and Operations Research*, 36(10):2773–2785, 2009b.

[9] S. Creemers and M.R. Lambrecht. Queueing models for appointment-driven systems. *Annals of Operations Research*, 178(1):155–172, 2010.

[10] C. S. Kim, I. Djamaludin, and D. N. P. Murthy. Warranty and discrete preventive maintenance. *Reliability Engineering & System Safety*, 84(3):301–309, 2004.

[11] A. Krontiris and G. Balzer. Failure distribution of repairable units approximation through markov chains. *16th Power Systems Computation Conference (PSCC 2008)*, 2008.

[12] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods and stochastic modeling*. SIAM, Philadelphia, 1999.

[13] H.L. Lee, V. Padmanabhan, and S. Whang. Information distortion in a supply chain: the bullwhip effect. *Management science*, pages 546–558, 1997.

[14] C. E. Love, Z. G. Zhang, M. A. Zitron, and R. Guo. A discrete semi-markov decision model to determine the optimal repair/replacement policy under general repairs. *European Journal of Operational Research*, 125(2):398–409, 2000.

[15] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach*. John Hopkins University Press, 1981.

[16] M.F. Neuts. *Structured Stochastic Matrices of M/G/1 type and their applications*. Marcel Dekker, Inc., New York and Basel, 1989.

[17] R. Pascual and J. H. Ortega. Optimal replacement and overhaul decisions with imperfect maintenance and warranty contracts. *Reliability Engineering & System Safety*, 91(2):241–248, 2006.

[18] F. R. Richards. Comments on the distribution of inventory position in a continuous-review (s,s) inventory system. *Operations Research*, 23(2):pp. 366–371, 1975.

[19] H.M. Taylor and S. Karlin. *An introduction to stochastic modeling*. Academic Press, 1998.

[20] H.C. Tijms. *A first course in stochastic models*. Wiley, 2003.

[21] D. D. W. Yao, M. L. Chaudhry, and J. G. C. Templeton. Analyzing the steady-state queue gix/g/1. *The Journal of the Operational Research Society*, 35(11):pp. 1027–1030, 1984.

[22] P. H. Zipkin. *Foundations of Inventory Management*. McGraw-Hill, New York, 2000.