The Final Order Problem for Repairable Spare Parts under Condemnation

John P.J. van Kooten
j.p.j.v.kooten@gmail.com

Tarkan Tan*
t.tan@tue.nl

Department of Technology Management, Technische Universiteit Eindhoven
P.O. Box 513, 5600 MB, Eindhoven, the Netherlands

April 2007

Abstract

We consider a manufacturer of complex machines that offers service contracts to her customers, committing herself to repair failed spare parts throughout a fixed service period. The suppliers of spare parts often discontinue the production of some parts as technology advances and ask the manufacturer to place a final order. We address the problem of determining final orders for such spare parts. The parts that we consider are repairable, but they are subject to the risk of condemnation. We build a transient Markovian model to represent the problem for a repairable spare part with a certain repair probability and repair lead time and we present some approximations that allow for further real life characteristics to be included. Furthermore, an approximate model that can be computed more efficiently is presented, and the sensitivity of the results obtained with respect to the problem parameters for both of the models is discussed.

Keywords: final order, repairable inventory theory, condemnation, spare parts management

1. Introduction and Motivation

In this paper we consider a manufacturer of complex machines that offers lasting service contracts to its customers. In these contracts the manufacturer commits herself to repair failed machines and provide spare parts for those machines at customers' sites throughout a fixed maximum contract service period. Especially for complex technological machines, customers' interest in service contracts is apparent. The maximum contract service period typically exceeds the production lifecycle of the machine by many years

* *Corresponding author*

due to advances in technology. The period after ending the production of a machine is the last phase of the product lifecycle and is often referred to as *end-of-life (service) period* or *final phase*. This final phase can give rise to unavoidable and undesirable issues for inventory management. Demand rates for spare parts are declining towards the end of the product lifecycle and an estimation of remaining demand for these parts is difficult to make due to technological improvements and external factors. Manufacturers are concerned with minimizing stock levels of spare parts towards the end of life of a product. Suppliers are also concerned with the management of this final phase with regard to their components. Suppliers may choose to discontinue the production of the spare part considered and ask a manufacturer to place a final order, or end-of-life buy, to cover for all demand in the remaining contract service period. As a consequence, the manufacturer is forced to make a possibly large final purchase of spare parts under uncertain future demand to form her spare parts pool, in order to be able to comply with the constraints as laid down in service contracts. This is the case even if the spare part is repairable, since successful repair is not always possible and repair lead times may be (too) long. Stocking large quantities of spare parts with uncertain future demand is unfavorable for a manufacturer of complex machines, since remaining inventory at the end of service period turns obsolete and is often scrapped. Cattani and Souza (2003) report that scrapping of obsolete inventory can reduce profits by up to 1% of revenue each year. The effect of inventory obsolescence is especially important for repairable items, since in general repairable items are relatively expensive and the number of times the part can be repaired (and thus used) is uncertain. In many supply systems for the maintenance of expensive high technology equipment, a large portion of investment is tied up in the inventory of repairable spare parts. For example, both Schrady (1967) and Sherbrooke (1971) report that repairable spare parts are responsible for over 50% of inventory investment within the U.S. Air force and U.S. Navy. The repairable spare parts, upon failure, are sent back to the supplier who attempts repair. Successful repairs are used for future replacements. Defective spare parts that cannot be repaired or for which repair is no longer economically feasible are condemned. In this paper we present a methodology to calculate final orders for repairable items, specifically taking the effect of condemnation into account.

We implemented this methodology at a manufacturer of complex machines for nanotechnology markets in Eindhoven, the Netherlands. The company produces several diagnosis and analysis tools which are used by research institutes and production companies worldwide, often in clean room environments. The company stocks several thousands of active spare parts and faces contract periods on systems of up to 25 years after production stop of the system. Repairable spare parts are responsible for over 60% of spare part inventory investment. Due to the high-technology nature of the company's products and long service periods, final ordering and condemnation of repairable parts are important issues; condemnation rates ranging from 5% to 30% typically. Besides final ordering, the company needs to adjust its estimate of excess spare parts stock on a quarterly basis for its obsolescence reserve. To do so, the company needs a methodology to calculate final orders as total demand estimate in the contract service period on a quarterly basis for all repairable spare parts in inventory.

The rest of this paper is organized as follows. In section 2, we review the related literature. In section 3, we present a transient Markovian model, specifically incorporating lead-time effects. Section 4 presents the analysis to obtain the optimal final order. In section 5 implementation issues are covered that were faced during the mentioned implementation. In section 6 an approximate binomial model is presented that considers the repairable equivalent of consumable demand. Numerical results comparing the Markovian model and the binomial model with simulation results are presented in section 7. Finally, in section 8, conclusions are drawn and some possible directions for future research are presented.

2. Related Literature

There exists a vast amount of literature on the management of spare parts inventories (both consumable and repairable). We refer the reader to Kennedy et al. (2002) for an excellent review. In this section however, we refine ourselves to literature that is more closely related to the problem of final ordering for spare parts in general and repairable items and condemnation more specific.
While the final ordering problem for consumable spare parts (parts that leave the system permanently after satisfying demand) is studied by a number of authors, the literature on the final ordering for repairable spare parts is limited. Although some work exists that considers closed loop inventory systems where repair is always successful, we are not aware of any work that considers the case of repairable items under condemnation. We distinguish the literature into work on consumable final order theory and repairable final order theory.

Literature on the final ordering problem for consumable spare parts at the beginning of a final phase is not abundant. Moore (1971) develops a method to forecast the 'all-time requirement' of consumable service parts in the motor-car industry. By plotting sales data on a logarithmic scale the author obtains three families of curves to be common for 85% of the spare parts considered. Ritchie and Wilcox (1977) develop a method to forecast all-time future demand for spare parts to time the moment of the final production run using renewal theory. Hill et al. (1999) treat the more general problem of determining stock replenishment policies for consumable spare parts for machines that are no longer manufactured. They use a dynamic programming approach to derive optimal policies and they propose a newsboy approach to determine the optimal replenishment size if there is just one opportunity to place a final order. Fortuin (1980, 1981) describes the calculation of final orders or all-time requirement for consumable spare parts. He assumes an exponentially decreasing demand pattern and uses normal approximation to derive expressions for several curves that indicate the size of the needed final order to attain a certain service level. Teunter and Fortuin (1999) and Teunter (1998) derive a dynamic programming algorithm for calculating minimal costs and the size of the corresponding final order. The authors include costs for initial provisioning, inventory holding, penalty costs per stockout and disposal cost and they present an approximation by means of a newsboy model. The authors include the possibility of re-supply through some external channel but assume that it occurs at no extra cost and thus the re-supply is only used to decrease the actual demand for spare parts. Teunter and Fortuin (1998) extend their work by presenting an application of their work at the service organization of a large electronics manufacturer in the Netherlands. Cattani and Souza (2003) study in their paper the effect of delaying a final purchase under several scenarios of remaining demand. They contrast the manufacturer's benefit with the cost incurred by the supplier to show that the supplier is likely to require an incentive to enact a delay of the final purchase.

Although there is no work specifically aimed at the final ordering problem for repairable parts under condemnation to the best of our knowledge, some general work that considers closed loop repairable systems yield expressions for the desired pool size of a repairable item assuming perfect repair. Gross et al. (1977) use queueing models to determine the optimal number of repair channels and spare parts needed to support a finite population of items which break down at random times and require repair. Walker (1996) presents a graphical aid to determine for a given number of machines, the needed repair pool to attain a certain service level constraint based on the ratio of the mean lead-time to the mean failure-free operating time. Rustenburg (2000) presents a multi-echelon approach to determine how many spare parts to buy and at what echelon to locate them in order to maximize the availability of the complex equipment supported by the repair pool.

Condemnation in general repairable inventory theory is studied extensively; however those works assume that it is always possible to repurchase spare parts from an external supplying source when needed. The work that is closest related to our problem is that of Inderfurth and Mukherjee (2006) who characterize three possible options for organization of spare part acquisition after stopping regular production. These options are a final order, performing extra production runs or remanufacturing to gain spare parts from used products. They model and solve the problem by a decision tree and a stochastic dynamic programming procedure. While remanufacturing is possible in their work, the authors do not specifically consider condemnation of repairable parts but treat the option of remanufacturing of returned (consumable) parts as source for future demand.

The combination of condemnation and the final order problem for repairable spare parts is to our knowledge not covered in literature yet. The main contribution this paper makes is developing a methodology for obtaining a final order for repairable spare parts taking condemnation specifically into account. Besides an exact solution, some approximations to the base Markovian model are presented that allow for further real life characteristics to be included. Furthermore, next to the Markovian model, a binomial model that can be computed more efficiently is presented; which is shown to work well for certain parameter ranges.

3. Model, Assumptions and Definitions

We consider a final ordering situation for a single spare part that does not interact with other parts. We model the problem as a continuous-time transient Markov chain, which we present in this section. Let q_{ab} denote the state transition rate from state a into state b. We denote the expected sojourn time in state a given that the transition is into state b as τ_{ab}. We denote the state space of the transient states as Ω_s and that of the absorbing states as Ω_r. The total space set is then defined by $\Omega = \Omega_s \cup \Omega_r$.

The following assumptions to the model are made:

1. Failures of a spare part occur due to a Poisson process with known parameter λ. In total there are m systems featuring the spare part; the total yearly demand rate thus becomes $m\lambda$. We assume that each system m features only one spare part. The assumption of Poisson demand is a reasonable assumption for demand for spare parts as discussed in general spare parts literature.
2. A defective spare part is immediately attempted to be repaired upon its arrival to a repair shop or supplier. A defective part can be repaired successfully with probability p and is condemned with probability $1 - p$. It may be the case that a repair attempt in reality can have more outcomes than assumed here (e.g. imperfect repair at a lower quality level instead of either successful or condemnation); however we assume that repair attempts which result in "less than perfect" quality are not desirable and thus are considered as condemnation. Furthermore, imperfect repair might lead to an increased failure rate of other components in the system, which we do not consider here.
3. Repair lead times are exponentially distributed with a known repair rate of μ parts per year per repair channel independent of the number of parts in repair (we assume that there are ample servers). This assumption can be justified by looking upon the repair lead times as delivery times instead of the throughput of a queueing system. We note that our model can easily be adapted to a situation where a finite number of servers is assumed. The repair leadtime consists of the time between replacement of the part to the moment the repaired part is either returned as available for use to the spare parts pool or condemned.
4. When a spare part fails, it is assumed to be replaced instantaneously, as long as there is a spare part available. Otherwise, the customer waits until a spare part becomes available from repair. The problem is modeled in this way, since in general the travel distance to a customer site and associated costs are large enough to justify a one-stop-and-exchange policy. Furthermore, due to technical reasons (e.g. dust, contamination or sanitation) it may be undesirable to work on a system in two separate occasions.
5. A one-for-one repair policy is used (e.g. we assume no batching of repairs). This assumption is justifiable for expensive parts; in general repairable parts are more expensive than consumable parts.

Let (i, j) denote a state where i represents the number of good parts on hand, with $i \geq 0$ representing an on- hand quantity and $i < 0$ a backorder situation; and j represents the number of defective parts with $0 \leq j \leq N$, the final order being N. We have $(i + j) \leq N$. Let $(N, 0)$ be the starting state of the system. From state $(N, 0)$ the system enters state $(N - 1, 1)$ with rate $m\lambda$ indicating a failure of a system and thus demand for a spare part. From this state three future states may be reached: $(N, 0)$ with rate μp indicating successful repair; $(N - 1, 1)$ with rate $\mu(1 - p)$ indicating condemnation of a part and $(N - 2, 2)$ with rate $m\lambda$ indicating another failure of a system and thus demand for a spare part before completion of the part in repair. The transient behavior of the system continues until one of so-called 'absorbing states' is reached. An absorbing state is defined as a state which, once entered, cannot be left and represents an undesirable situation. We distinguish two absorbing states in our model. Firstly, the absorbing state $(\gamma, 0)$, $\gamma < 0$, represents a situation where there are no parts on hand or in repair and there is (backlogged) demand for a spare part which implies that service (in terms of supplying repaired parts) is no longer possible. Note that the state (0, 0) is not considered an absorbing state in our model; only upon occurrence of demand in that state the system reaches absorption through $(\gamma, 0)$. Secondly, the absorbing state (NS), short for (*No Service*), represents an absorbing state that is introduced to limit the amount of backorders in the system. Without a bound on the number of backorders, an extremely unfavorable situation can be reached in terms of service (e.g. probability of eventually being able to supply all backordered demand can become undesirably small or the time until delivery of backlogged demand can become undesirably long). Therefore we let B denote the maximum number of allowed backorders $(B \geq 0)$. B is assumed to be a constant number independent of the number of parts in repair and reflects the above mentioned issues like delivery time and likelihood of fulfilling all backordered demand. The state (NS) can be reached in case any demand occurs and the model is in any of the maximum backordering states $(-B, j)$. Determining B is an important decision variable since it determines the state space of the model, which we will discuss later. Once the model enters a backordering state, the customers' systems are not operational and they have to wait for a part to return from repair as available for use. Note that modeling into two separate absorbing states is only for illustrative purposes; modeling the system with one absorbing state combining the rates into (NS) and $(\gamma, 0)$ is also possible.

When the probability of absorption in (NS) is high, the system will spend a relatively large amount of time in a backordering state and it is likely that there will be some parts remaining in the repair pool at the end. On the other hand, if the absorption probability of $(\gamma, 0)$ is relatively high, this indicates the system is able to "cope" with the incoming flow of defective parts in general. It is in those instances that the system is likely to end up with a very limited number of parts remaining in the repair pool. From a practitioners viewpoint this last situation is favorable because there is no large write-off of repairable parts due to absence of demand; the parts that are written-off are done so due to condemnation throughout the contract service period.

Having defined the state space and the behavior of the model, we can distinguish the possible transitions and the according rates into seven categories, see Table 1:

condition	from state	to state	event	rate
$i > 0, j = 0$	$(i, 0)$	$(i-1, 1)$	failure	$m\lambda$
$i \geq 0, 1 < j \leq N$	(i, j)	$(i-1, j+1)$ $(i+1, j-1)$ $(i, j-1)$	failure successful repair condemnation	$m\lambda$ $j\mu p$ $j\mu(1-p)$
$-B < i < 0, 1 < j \leq N$	(i, j)	$(i-1, j)$ $(i+1, j)$ $(i, j-1)$	failure successful repair condemnation	$(m+i)\lambda$ $j\mu p$ $j\mu(1-p)$
$i = -B, 1 < j \leq N$	$(-B, j)$	(NS) $(i-(i-1+B), j)$ $(i-(i+B), j-1)$	failure, absorption successful repair condemnation	$(m-B)\lambda$ $j\mu p$ $j\mu(1-p)$
$-B < i < 0, j = 1$	$(i, 1)$	$(i-1, 1)$ $(i+1, 0)$ $(\gamma, 0)$	failure successful repair condemnation, absorption	$(m+i)\lambda$ μp $\mu(1-p)$
$i = -B, j = 1$	$(-B, 1)$	(NS) $(i-(i-1+B), j)$ $(\gamma, 0)$	failure, absorption successful repair condemnation, absorption	$(m-B)\lambda$ μp $\mu(1-p)$
$i = 0, j = 0$	$(0, 0)$	$(\gamma, 0)$	failure, absorption	$m\lambda$

Table 1, Transaction Types and Corresponding Rates in the Markov Chain

Considering the above-mentioned assumptions, the Markov chain can now be composed with starting state $(N, 0)$, see Figure 1. Note that as a result of our assumption that failed parts will only be replaced upon availability of a part, the maximum repair rate becomes $N\mu$ while there may be up to $N + B$ failed parts in the system. Furthermore, observe that the total repair rate $N\mu$ is divided in two separate rates; due to the 'thinning-property' this results in two separate Poisson processes with rates $N\mu p$ and $N\mu(1-p)$. Due to the assumption of instantaneous repair, the total demand rate equals $m\lambda$ as long as there is no backordered demand; in case demand is backlogged the total demand rate becomes smaller than $m\lambda$ since non-operating systems cannot generate demand for the spare part until they are operational again.

We are interested in finding the distribution of time until absorption into one of the defined absorbing states. Our purpose is to find the optimal value of N using this Markov chain such that a pre-specified contract service period, CSP, for the spare part is met with a specified probability. The time until absorption into one of the two absorbing states is defined as the remaining service period given a repair pool of size N, RSP_N. Observe that a larger value of N will result in a larger value of RSP_N. The final order N^* is selected such that RSP_{N^*} is greater than or equal to CSP with a specified probability. This specified probability is referred to as the "service level", SL, and measures the probability of covering all arrivals in the given CSP.

$$N^* = Min\{N : \text{Prob}\{RSP_N \geq CSP\} \geq SL\}$$

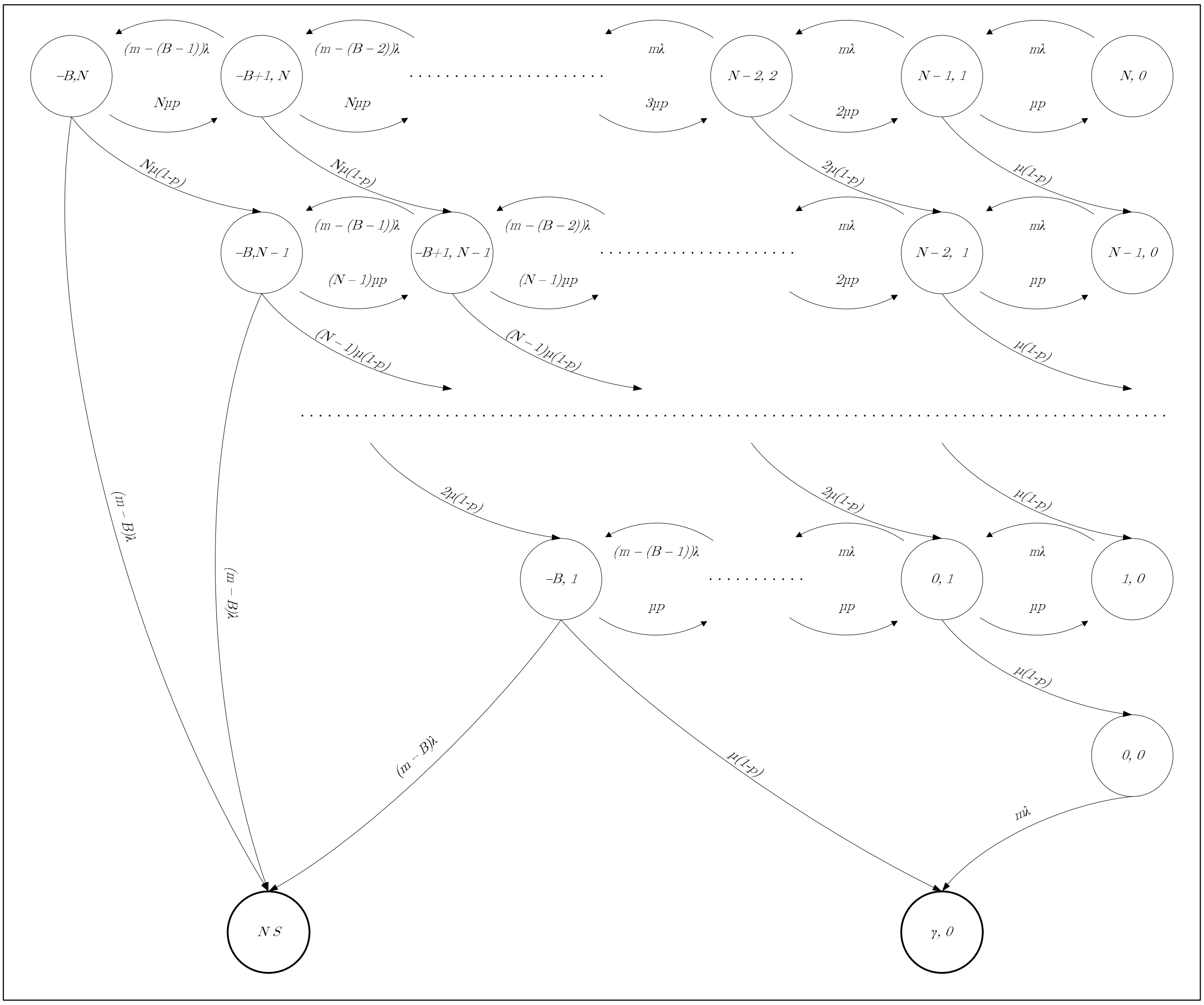

Figure 1, Markov Chain

We already mentioned that determining B is important in defining the state space of our model. We now propose two methods to set B. Firstly, B may be set by defining a maximum time that is allowed to pass until a backorder is delivered, $T_{deliver}$. The number of maximum backorders that in this case may occur can be obtained by multiplying the maximum repair rate $(N \cdot \mu)$ by $T_{deliver}$, yielding the selection rule

$$Max\{B : B \leq T_{deliver} \cdot N \cdot \mu\}$$

The B obtained in this way depends only on time. Another selection rule for B is dependent on the possibility of fulfilling all backlogged demand. A decision maker may set a desired probability $p_{deliver}$ such that the probability of all of the B parts in backlog being repaired successfully, $p^{[B]}$, is greater than or equal to $p_{deliver}$. The probability $p^{[B]}$ can be calculated as follows: In case $B \geq N$, all B repairs must be successful; in case $B < N$, B or more repairs out of N parts in repair must be successful. This leads to the following expression:

$$p^{[B]} = \begin{cases} p^{(N+B)} & \text{for } B \geq N \\ \sum_{k=B}^{N} \binom{N}{k} p^k (1-p)^{N-k} & \text{for } B < N \end{cases}$$

The largest value of B that still satisfies $p^{[B]} \geq p_{deliver}$ can be selected as the maximum number of allowed backorders. These bounds on B using $T_{deliver}$ and $p_{deliver}$ may be used to obtain feasible values for the situation specific circumstances. Although we consider a fixed B in the Markovian model, a dynamic backorder level that depends on the size of the actual repair pool could also be used.

4. Analysis

We are interested in calculating the time until one of the absorbing states is reached, considering the starting state. The analysis of this section follows the standard analysis of transient discrete Markov chains. The reader is referred to Kemeny and Snell (1976) for an in-depth discussion and proof of the transient Markovian analysis presented here.

To calculate the time until absorption, we translate the continuous-time Markov process into a discrete-time Markov chain. Let p_{ab} denote the state transition probability from state a into state b and let P denote the matrix of state transition probabilities $\{p_{ab}\}$. From transition rates, transition probabilities can be calculated as

$$p_{ab} = \frac{q_{ab}}{\sum_d q_{ad}} = \frac{q_{ab}}{q_a} \quad \text{for } a, b \in \Omega,\ a \neq b \text{ and } p_{aa} = 0$$

P can be rearranged into the canonical form which divides the matrix into four sub-matrices as depicted below:

$$P = \begin{array}{c} \overbrace{\quad}^{s} \; \overbrace{\quad}^{r} \\ \left[\begin{array}{c|c} Q & R \\ \hline O & I_r \end{array}\right] \end{array} \begin{array}{l} \} \, s \\ \} \, r \end{array}$$

Within P, Q is an (s x s) matrix stating transitions to transient states; R is an ($s \times r$) matrix stating transitions from transient to absorbing states; O is an ($r \times s$) matrix consisting of zeroes and I_r is an ($r \times r$) identity matrix. Note that we have $r = 2$ and s depends on B and N. For $a \in \Omega_s$, define $\tau_a = \tau_{ab}$ with probability p_{ab}, then the k^{th} moment of sojourn time in a state a, $E[\tau_a^k]$, is defined as

$$E\left[\tau_a^k\right] = \sum_{b=1}^{r} p_{ab} E\left[\tau_{ab}^k\right], \quad k = 1,2,... \qquad \text{and} \qquad E\left[\tau_{ab}^1\right] = \frac{1}{q_a}$$

Let M be the diagonal matrix $\left(E[\tau_1], E[\tau_2], ..., E[\tau_s]\right)$ and I be the (s x s) identity matrix. For $a, b \in \tilde{\Omega}$ define L_{ab} as the total time in state b before absorption, given starting state a. Then $L = \{E[L_{ab}]\} = (I - Q)^{-1} M$ is the (s x s) matrix giving the total expected time in transient states. The mean times to absorption, provided that the process started in any of the Ω_s states, are given by $V^{(1)}$:

$$V^{(1)} = L\xi$$

with ξ a column vector of length s consisting entirely of ones. The 1st element of $V^{(1)}$ gives the first moment of time to absorption for RSP_N, denoted by μ_N. Let furthermore $R^{(2)}$ be a column vector $\left(E\left[\tau_1^2\right], E\left[\tau_2^2\right], ..., E\left[\tau_s^2\right]\right)^T$ (where α^T stands for α transposed) and for $a, b \in \Omega_s$ let M' be the matrix $\{p_{ab} E[\tau_{ab}]\}$. The second moment of time to absorption, given the process started in any of the Ω_s states, is denoted by $V^{(2)}$:

$$V^{(2)} = (I - Q)^{-1}\left[R^{(2)} + 2M'V^{(1)}\right]$$

The 1st element of $V^{(2)}$ gives the second moment of time to absorption for RSP_N, from which the standard deviation of the time to absorption, σ_N, can be calculated. The probability of ending up in absorbing state r $(r \in \Omega_r)$, given the process started in transient state s $(s \in \Omega_s)$ is given by $p_{abs}(r)$. These probabilities are given in W, the resulting matrix of multiplication of L and R:

$$W = LR$$

The distribution of the time to absorption is of phase-type. From the known first and second moment, an approximating distribution can be obtained. Let c_N be the coefficient of variation of time to absorption. In case $0 < c_N < 1$, a mixture of two Erlang distributions with scale parameters *k-1* and *k*, $\text{Erlang}_{k-1,k}$, is a common approximation; and in case $c_N \geq 1$ a hyperexponential distribution can be used (Tijms, 1994).

Finally, we note that the *SL* service measure is very conservative, since it is concerned with the probability that *all* arrivals are met. Therefore we define the "actual" service level, *ASL,* that the customers perceive, as the percentage of failures where the failed spare part is replaced during the *CSP*, either directly from the available pool or after some time when a repaired part has become available for use. Under the assumption that the arrival rate remains unchanged after the servicing stops, this service level is equal to the percentage of the time that spare parts are delivered during the *CSP*, which can be formulated as

$$ASL = \int_0^{CSP} \frac{rsp_N}{CSP} dF(rsp_N) + \Pr\{RSP_N \geq CSP\}$$

where $F(rsp_N)$ is the distribution function of RSP_N. In case this assumption is significantly violated, the above expression becomes an approximation. Under ASL as the service measure, the final order N^* should be selected as the smallest N which assures that ASL is greater than or equal to a pre-specified level.

5. Implementation Issues

In what follows we discuss some implementation issues which we faced during our earlier mentioned application.

5.1 Including changes in demand rate

The assumption of a constant installed base size, m, and failure rate, λ, may be weak assumptions considering real life characteristics. Typically, towards the end of the final phase, more and more customers will replace their machines with newer, more advanced machines. On the other hand, due to aging effects, failure rates of spare parts might increase over time. To overcome these shortcomings in the Markovian model, we present an approximation procedure here that uses a weighting factor to incorporate demand changes in the remaining service period of the spare part. In this approximation we consider discrete time periods. We let m_t denote the number of systems featuring the spare part in period t, $t = 1, 2, \ldots$ and let m_0 represent the current number of systems. The failure rate of the part per system in period t is denoted by λ_t, with λ_0 representing the current failure rate. The weighting factor, wf, weighs the demand based on failure rate and number of systems in CSP (denoted by λ_{CSP}, and m_{CSP} respectively) and can be calculated as

$$wf = \frac{m_{CSP}\lambda_{CSP}}{m_0\lambda_0}, \text{ where } \lambda_{CSP} = \frac{\sum_{t=1}^{t=CSP}\lambda_t m_t}{\sum_{t=1}^{t=CSP} m_t} \text{ and } m_{CSP} = \frac{\sum_{t=1}^{t=CSP} m_t}{CSP}.$$

Thus, instead of using the initial demand rate, a weighted average demand rate is used in this approximation. By introducing wf, in fact an adjusted but constant demand rate is assumed throughout the contract service period which is a weighted average of non-stationary demand rates. In reality, the interval between demand occurrences changes during the CSP. In this approximation we assume a constant interval between demands throughout the CSP which causes the mean value of time to absorption to remain unchanged. The variance term however will become different and although it still can be computed, it will not be the same as the actual variance term, as observed during our application. As a result the obtained pool size may be inflated by adding an "inflation factor" to obtain the desired SL. Since towards the end of the contract service period typically a small number of operating systems are expected to be actively in use, inflating the repair pool size with a very limited number of parts already leads to a relatively large improvement in SL. We obtained the inflation factor by means of simulation. During simulation in our application we obtained values for $wf < 0.50$ and corresponding inflation factors between 5% and 30% of μ_N, depending on μ_N, to obtain the desired confidence level set by the company decision makers. In this application we asked decision makers to indicate the expected behavior of m and λ during the remaining service period and obtained m_t and λ_t accordingly. Instead of using one expected behavior of m and λ, one may use several scenarios to obtain a weighted expected behavior based on the likelihood of any of the scenarios.

5.2 An alternative solution using regression

Although the matrix $(I - Q)$ is block upper-triangular, which is therefore relatively easier to invert, the computation of $V^{(1)}$ and $V^{(2)}$ can still take a significant amount of time due to the matrix inversion operation, as repair pool size and therewith the state space increases. Nevertheless, during a number of test runs we observed that μ_N can be approximated well as a linear function of N as long as $N > 5$, where the intercept and the slope of the linear relationship depend on the problem parameters. Accordingly, μ_N for large pool sizes can be calculated by extrapolating those of the smaller ones, which are computed as explained in Section 4. However, σ_N exhibits a concave increasing pattern and this needs to be taken into account in computing N^*.

6. An Approximate Model

As discussed in Section 5.2, the matrix inversion operation may become prohibitive in applying the proposed method for large pool sizes. In this section we present a simple approximate solution method that may be used for computing N^*, which also enables inclusion of some characteristics such as changing demand patterns.

Especially for large pool sizes (e.g. multiples of mean annual demand) the decision on the size of the final order can be based mainly on the expected number of parts that are condemned during the remaining contract service period. Due to the size of the needed repair pool in that case, repair lead time will become an issue only towards the end of the contract service period, since there will be sufficient spare parts in the pool until then. Using this observation, one may consider a surrogate model that considers consumable spare parts that stand for the "consumable equivalent" of repairable spare parts. For such a model, the number of times a spare part is expected to be used until condemnation needs to be made use of, together with the expected total number of failures during *CSP*. Let r denote the expected number of times a repairable part can be used, until further repair proves to be impossible or economically infeasible. A part can be used as new, and repaired n times successfully with probability p^n, resulting in the geometric series $r = 1 + p^1 + p^2 + p^3 + p^4 + \ldots$ if an infinite number of repair attempts is assumed. This series is equal to

$$r = \sum_{n=0}^{\infty} p^n = \frac{1}{1-p}.$$

Note that this is an approximation for r, since the maximum number of repairs will actually be limited as long as *CSP* is finite. Consequently, if N_c is the needed "consumable" spare pool size, then

$\tilde{N}^* = N_c / r = N_c(1-p)$ rounded up becomes the needed spare pool size in the repairable model, since the repair lead time issue is ignored.

We now discuss how to find N_c, using a periodic model. During our application, the available approaches in literature did not reflect the reality at the company very well (e.g. not all demand patterns were declining over time due to the possibility of increasing failure rates per spare part). As a result we present a binomial model for consumable demand, similar to the model of Kelle & Silver (1989) which is developed for reusable items (e.g. containers or crates) in the field of reverse logistics. We define a probability, p_t, that a spare part in one system out of installed base of m_t systems will fail in period t. The values of m_t and p_t may be estimated based on the available information and expert opinion by the decision makers. For example, the current failure probability p_0 could be estimated as d_0 / m_0, where d_0 denotes (possibly weighted) average of the number of observed failures originating from all systems in the installed base in the periods prior to the final order. Having defined the probability p_t, each part that is installed in a system may produce two outcomes in a (future) period t: failure (and thus demand for a spare part) or not. This can be seen as a Bernoulli experiment with probability p_t of 'failure' and $1 - p_t$ of 'continuing operation' and m_t the number of trials (systems in the installed base) containing the part in period t. Then, the number of failures in each period t has a binomial distribution with parameters m_t and p_t. The length of a period should be selected in such a way that p_t is reasonably small; otherwise the binomial model's accuracy deteriorates, since the part can fail at most once in a period. The resulting distribution of demand in the entire contract service period, D_{CSP}, assuming independent demands between periods, has an approximate normal distribution with mean μ_{CSP} and variance σ^2_{CSP} due to central limit theorem, where

$$\mu_{CSP} = \sum_{t=1}^{CSP} m_t p_t \qquad \text{and} \qquad \sigma^2_{CSP} = \sum_{t=1}^{CSP} m_t p_t (1 - p_t)\,.$$

Then, the pool size that will cover all of the demand in CSP with probability SL can be calculated as $N_c = \mu_{CSP} + \Phi^{-1}(SL)\sigma_{CSP}$, where Φ denotes the distribution function of standard normal distribution. Similar to the discussion in Section 4, if ASL is used as the service measure, then

$$ASL = \Pr\{D_{CSP} \le N_c\} + \int_{N_c}^{\infty} \frac{N_c}{D_{CSP}} dF(d_{csp}) = \Phi\left(\frac{N_c - \mu_{CSP}}{\sigma_{CSP}}\right) + N_c \cdot E\left[\frac{1}{D_{CSP}} \middle| D_{CSP} \ge N_c\right],$$

where $F(d_{csp})$ is the distribution function of D_{CSP}, which is Normal(μ_{CSP}, σ^2_{CSP}). The last expectation term above can be calculated numerically, and the smallest N_c should be selected which assures that ASL is greater than or equal to a pre-specified level.

In what follows we demonstrate that errors in estimation of the parameters m_t and p_t through the history of failures have a limited effect on the results of this approximate model. It may be possible that the information as to the size of the installed base is not accurate. Assume that the actual number of systems in the installed base is $\hat{m}_0 = n \cdot m_0$, where $n > 1$ (we consider the case that $\hat{m}_t \ge m_t$ because during our application records of the systems that actually failed were available and thus the actual installed base could only be larger than the recorded installed base). Then, considering the case that p_0 is estimated by d_0 / m_0, the actual current failure probability, $\hat{p}_0$, becomes:

$$\hat{p}_0 = \frac{d_0}{n \cdot m_0} = \frac{1}{n} p_0 .$$

This means that in case the total number of installed bases is underestimated, then the failure probability is overestimated. But then, the actual expected number of failures in the current period, $\hat{\mu}_0$, is

$$\hat{\mu}_0 = \hat{m}_0 \hat{p}_0 = n \cdot m_0 \frac{1}{n} p_0 = m_0 p_0 = \mu_0 ,$$

which is the same as the expected number based on the incorrect parameters m_0 and p_0. There is a difference in the variance term though. The actual variance of the number of failures in the current period, $\hat{\sigma}_0^2$, is

$$\hat{\sigma}_0^2 = \hat{m}_0 \hat{p}_0 (1 - \hat{p}_0) = n \cdot m_0 \frac{1}{n} p_0 (1 - \frac{1}{n} p_0) = \sigma_0^2 + \frac{n-1}{n} m_0 p_0^2$$

which is positive for $n > 1$ (and negative if $n < 1$). This means that in case the actual number of active systems in the installed base is underestimated, then the variance of the number of failures is also underestimated. As a consequence, failing to recognize this underestimation will result in a lower final order than needed to meet the specified service measure (and thus to an increased likelihood of stock outs at the end of the *CSP*). In case m_t is overestimated –due to systems that are not active anymore– variance is also overestimated. Although the mean value is again the same, the calculation based on the estimate is on the conservative side this time. In case of an overestimation of m_t, the condition $p_t < n$ must hold, so that $\hat{p}_t \leq 1$.

7. Numerical Results

To indicate the influence of the model parameters on the resulting final order values, we performed a numerical comparison between the final order sizes that are obtained by the Markovian model (N) and the approximate model ($\tilde{N}$) on one hand, and the optimal one obtained through simulation (N_{sim}) on the other hand. The results are presented in Table . We used initial demand rates $m\lambda$ of 2 and 5 (where we kept λ constant at 0.02 and changed m accordingly); p values of 90% and 75%, μ values of 2, 10 and 25 and B values of 1 and 4. Furthermore we used a remaining service periods of 10 years and also calculated the long run absorption probabilities in either of the two absorbing states. In Table , one parameter is changed at a time in each instance and three actual service levels of 90%, 95% and 99% are presented. We present the smallest pool size that results in an *ASL* that is greater than or equal to the presented value. The selected parameter values cover a wide range of feasible parameters as found during our application.

parameters, RSP 10 yrs																	
$m\lambda$	p	μ	B	ASL	N_{sim}	N	$\tilde{N}$	ASL	N_{sim}	N	$\tilde{N}$	ASL	N_{sim}	N	$\tilde{N}$	$p(\gamma,0)$	$p(NS)$
2	90%	2	1	99%	5	7	3	95%	4	5	3	90%	4	5	2	11%	89%
2	90%	2	4	99%	5	6	3	95%	4	4	3	90%	3	4	2	97%	3%
2	90%	10	1	99%	5	6	3	95%	4	4	3	90%	3	3	2	77%	23%
2	90%	10	4	99%	5	5	3	95%	4	4	3	90%	3	3	2	100%	0%
2	90%	25	1	99%	5	5	3	95%	4	4	3	90%	3	3	2	95%	5%
2	90%	25	4	99%	5	5	3	95%	4	4	3	90%	3	3	2	100%	0%
2	75%	2	1	99%	8	10	7	95%	7	8	6	90%	6	7	5	29%	71%
2	75%	2	4	99%	8	9	7	95%	7	7	6	90%	6	6	5	79%	21%
2	75%	10	1	99%	8	9	7	95%	7	7	6	90%	6	6	5	89%	11%
2	75%	10	4	99%	8	9	7	95%	7	7	6	90%	6	6	5	100%	0%
2	75%	25	1	99%	8	9	7	95%	7	7	6	90%	6	6	5	98%	2%
2	75%	25	4	99%	8	9	7	95%	7	7	6	90%	6	6	5	100%	0%
5	90%	2	1	99%	10	13	6	95%	9	11	5	90%	8	10	5	0%	100%
5	90%	2	4	99%	10	11	6	95%	8	9	5	90%	7	8	5	6%	94%
5	90%	10	1	99%	9	10	6	95%	7	8	5	90%	6	7	5	37%	63%
5	90%	10	4	99%	9	9	6	95%	7	7	5	90%	6	6	5	93%	7%
5	90%	25	1	99%	9	10	6	95%	7	8	5	90%	6	7	5	77%	23%
5	90%	25	4	99%	9	9	6	95%	7	7	5	90%	6	6	5	100%	0%
5	75%	2	1	99%	18	20	14	95%	15	17	13	90%	14	15	12	3%	97%
5	75%	2	4	99%	17	18	14	95%	14	15	13	90%	13	13	12	21%	79%
5	75%	10	1	99%	17	18	14	95%	14	15	13	90%	13	13	12	61%	39%
5	75%	10	4	99%	17	18	14	95%	14	15	13	90%	13	13	12	97%	3%
5	75%	25	1	99%	17	18	14	95%	14	15	13	90%	13	13	12	89%	11%
5	75%	25	4	99%	17	18	14	95%	14	15	13	90%	13	13	12	100%	0%

Table 2, Comparison of Markovian model and Approximate model, RSP = 10 years

For the optimal solution of the problem, on top of the self-explanatory results, we observe the following:

1. Adding relatively few spare parts to the pool improves *ASL* significantly, due to relatively high repair probabilities.
2. Improving repair probability is more favorable compared to reduction in repair leadtimes, since a higher repair probability has economically more value for the parameter range that we consider.
3. To attain a certain service level as the demand rate increases, the rate of increase in the required pool size is less than that of the demand rate, due to variability pooling.
4. The optimal pool size is sensitive to changes in B only when the system intensity is high.

We observe that in most cases N is close to N_{sim}, and in all cases $N \geq N_{sim}$. The Markovian model is over-sensitive to reduction in B and it performs relatively poor in that case, as well as the case where there is a high absorption probability of the state *(NS)*. The latter case indicates that the system is slow in repairing parts compared to the arrival of new defective parts, and therefore significantly more parts are needed in that case due to long repair lead times. Furthermore, for lower B values absorption through *(NS)* is more likely. The difference between the results of these two models in those instances can be explained by two reasons. First of all, the coefficient of variation is higher for lower B values and higher *(NS)* probabilities, which deteriorates the performance of the Erlang approximation. Secondly, we assume in our definition of *ASL* that the arrival rate remains unchanged after the servicing stops, whereas the arrival rate will actually become smaller as machines malfunction and not turn operational again. Nevertheless, we do not claim that $N \geq N_{sim}$ necessarily needs to be the case for all possible problem instances; because all arrivals before absorption or before the end of *CSP* are assumed to be met for the computation of *ASL* in the Markovian model, which does not necessarily hold.

As for the approximate (binomial) model, we first note that its performance is in general poorer than that of the Markovian model, as expected, since the calculated final order does not depend on μ and B. Consequently, the binomial model performs poorly in systems with higher intensity -i.e. when the arrival rate is significantly larger than the repair rate-, because it only accounts for condemnation of non-repairable parts and does not consider repair lead time effects, and in systems where higher B values are not allowed. We also observe that that $\tilde{N} < N_{sim}$ in all cases considered, and in general the difference N_{sim} - $\tilde{N}$ increases as *ASL* increases. This is mainly because of the underestimation of the variability by the normal approximation.

Since the pool size obtained by using the Markovian model appears to serve as an upper bound to the optimal pool size, and that by the approximate model as a lower bound, one might use a correction factor that inflates or deflates the pool sizes to reach the desired service level; such as a correction factor (for the approximate model) of +1 for ASL = 90% and +2 for ASL = 95% for the results that we presented in Table 2. Another possibility is to use a weighted average of the two, where the weights depend on the problem parameters. In the results that we present in Table 2, even a simple average (rounded up) finds the optimal pool size in all of the 24 cases for ASL=90%, and the difference is never more than 1 in any of the 72 cases considered.

8. Conclusions and Future Work

In this paper we build a model which makes it possible to calculate final orders for a repairable spare part considering a predefined service level and explicitly taking condemnation into account. These final orders typically occur during the final phase of the lifecycle of the product that is supported by the spare part. The problem is modeled as a transient Markov chain to calculate the first and second moment of the time until absorption, resulting in an approximate distribution of the time until absorption. Accordingly, a final order size is obtained that guarantees a certain service level during the contract service period. We find out that a linear relationship between the pool size and the remaining service period exists for relatively large pool sizes. This property is useful, considering significant computer effort that is required to solve the Markovian model for larger repair pool sizes. Furthermore, we discuss how to incorporate some real-life characteristics we encountered during the implementation of this methodology at a manufacturer of complex technological machines in the Netherlands. We also present an approximate model that calculates a final order based on "consumable" equivalent demand for the spare parts, considering the expected number of possible repairs before the part is condemned. The approximate model requires limited computer effort and we include several real-life characteristics in this approximate model, as well. We discuss the sensitivity of the results obtained with respect to the problem parameters for both of the models.

Our work could be extended in several ways. An extension that considers batching of repair jobs could be useful for some environments. The model could also be extended by a cost approach that includes the possibility of another order later at a high cost. Finally, considering different customer classes would be an interesting extension, where the classification may be due to different service contracts, criticality of the customer, etc. In such a system the customers that belong to certain classes would not be served in certain states of the system, so that a reservation for higher class customers takes place.

References

Cattani, K.D., and Souza, G.C. (2003), "Good buy? Delaying End of Life Purchases", *European Journal of Operational Research*, 146, p. 216 – 228.

Fortuin, L. (1980), "The All-Time Requirement of Spare Parts for Service after Sales – Theoretical Analysis and Practical Results", *International Journal of Operations and Production Management* 1(1): p. 59 – 69

Fortuin, L. (1981), "Reduction of the All-Time Requirement for Spare Parts", *International Journal of Operations and Production Management* 2: p. 29 – 37

Gross, D., Kahn, H.D., and Marsh, J.D. (1977), "Queueing Models For Spares Provisioning", *Naval Research Logistics Quarterly*, 24, p. 521 – 536

Hill, R.M., Omar, M., and Smith, D.K. (1999), "Stock replenishment policies for a stochastic exponentially-declining demand process", *International Journal of Operational Research* 116, p. 374 – 388

Inderfurth, K., and Mukherjee, K. (2006), "Analysis of Spare Part Acquisition in Post Product Lifecycle", *Working Paper No.6,* Faculty of Economics and Management Magdeburg, Otto von Guericke University, Magdeburg

Kelle, P., and Silver, E.A. (1989), "Forecasting the Returns of Reusable Containers", *Journal of Operations Management*, 8 (1), p. 17 – 35

Kemeny, J.G., and Snell, J.L. (1976), "Finite Markov Chains", Springer Verlag, Berlin

Kennedy, W.J., Wayne Patterson, J., and Fredendall, L.D. (2002), "An Overview of Recent Literature on Spare Parts Inventories", *International Journal of Production Economics*, 76, p. 201 – 215.

Moore, J.R. (1971), "Forecasting and Scheduling for Past-Model Replacement Parts", *Management Science,* Vol 18, No 4, p. B-200 – B-213

Ritchie, E., and Wilcox, P. (1977), "Renewal theory forecasting for stock control", *European Journal of Operations Research*, Vol 1, p. 90 – 93

Rustenburg, J.W. (2000), "A System Approach to Budget-Constrained Spare Parts Management", PhD Thesis, Technische Universiteit Eindhoven

Schrady, D.A. (1967), "A Deterministic Inventory Model for Repairable Items", *Naval Research Logistics Quarterly*, 14, p. 391 – 398

Sherbrooke, C.C. (1971), "An Evaluation for the Number of Operational Ready Aircraft in Multi-Level Supply Systems", *Operations Research*, 19, p. 618 – 635

Teunter, R.H., and Fortuin, L. (1998), "End-of-life service: A case study", *European Journal of Operational Research*, 107, p. 19 – 34

Teunter, R.H., and Fortuin, L. (1999), "End-of-life service", *International Journal of Production Economics*, 59, p. 487 – 497

Teunter, R.H. (1998), "Inventory Control of Service Parts in their Final Phase", Ph.D. Thesis, Graduate School/Research Institute of Systems, Organisation and Management, University of Groningen, The Netherlands

Tijms, H.C. (1994), "Stochastic Models, an Algorithmic Approach", John Wiley & Sons, Chichester

Walker, J. (1996), "A Graphical Aid for the Initial Purchase of 'Insurance Type' Spares", *Journal of the Operational Research Society*, Vol. 47, p. 1296 – 1300