Int. J. Production Economics 118 (2009) 260–268

Contents lists available at ScienceDirect

Int. J. Production Economics

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

Joint optimisation of spare part inventory, maintenance frequency and repair capacity for *k*-out-of-*N* systems

Karin S. de Smidt-Destombes [a,*], Matthieu C. van der Heijden [b], Aart van Harten [b]

[a] TNO Defence, Security and Safety, The Hague, The Netherlands
[b] University of Twente, Enschede, The Netherlands

ARTICLE INFO

Available online 10 October 2008

Keywords:
Maintenance
Spare parts
Repair capacity
k-Out-of-*N* systems
Availability
Installed base

ABSTRACT

To achieve a high system availability at minimal costs, relevant decisions include the choice of preventive maintenance frequency, spare part inventory levels and spare part repair capacity. We develop heuristics for the joint optimisation of these variables for (a) a single *k*-out-of-*N* system under condition-based maintenance and (b) an installed base of multiple identical *k*-out-of-*N* systems under block replacement. We show that a straightforward extension of the METRIC method for spare part inventory optimisation yields inferior results, because both the availability and costs are not necessarily monotonous functions of the decision variables. We develop an adjusted marginal analysis and show that it performs considerably better in numerical experiments.

1. Introduction

Many of today's capital assets require a high availability, because the consequences of downtime can be serious. For example, the failure of a wafer stepper in the semiconductor industry usually leads to a production stop. This yields reduced output and hence reduced revenues, so the consequence of a failure is serious. The system availability is influenced by many tactical and operational decisions, such as the maintenance frequency, the amount of maintenance resources like service engineers and test equipment, and spare part inventories. A common approach is to decompose the overall trade-off in a set of subproblems. However, we can argue that there are clear relations between these decision variables.

For example, consider the interaction between spare part inventories and maintenance frequency. Demand for spare parts arises from both preventive and corrective maintenance. A higher preventive maintenance frequency leads to higher maintenance cost, but at the same time to a better predictable demand for spare parts and hence to a lower spare parts safety stock. Also, the interaction between repairable spare part inventories and the capacity needed to repair these spares cannot be neglected, see e.g. Sleptchenko et al. (2002, 2003): low repair shop capacity means a high utilisation, so long spare part repair leadtimes. As safety stocks should cover the demand during the leadtime, this means that savings on repair capacity lead to a need for more spare parts and vice versa.

In this paper, we discuss heuristics for the joint optimisation of maintenance frequency, spare part inventories, and spare part repair capacity. We focus on *k*-out-of-*N* systems with hot stand-by redundancy. That is, a system consists of *N* identical components of which only $k<N$ are required for system operation. The $N-k$ stand-by components have the same failure behaviour as the *k* operational components. We construct our optimisation heuristics based on approximations that we have developed before to calculate the system availability as function of the maintenance frequency, spare part inventories and repair capacity (De Smidt-Destombes et al., 2004, 2006, 2007). A complication is that the availability might not be a monotonous function of the maintenance

* Corresponding author.
E-mail address: ksmidt@feweb.vu.nl (K.S. de Smidt-Destombes).

0925-5273/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpe.2008.08.058

frequency. When the frequency decreases, the probability that the system fails before maintenance starts increases and this pushes the availability down. On the other hand, the cycle length increases and the expected uptime in a cycle increases as well, which pushes the availability up. The aggregate effect may both be a decrease or an increase in the system availability. Therefore, the development of a joint optimisation method for spare part inventories, repair capacity and maintenance frequency is not straightforward.

The remainder of this paper is structured as follows. In the next section, we discuss the related literature. In Section 3, we develop an optimisation heuristic for a single k-out-of-N system with condition-based maintenance. We evaluate the quality of our heuristic in a numerical experiment in Section 4. We give our conclusions and directions for further research in Section 5.

2. Related literature

Although there is a lot of literature on spare part management (e.g. Sherbrooke, 2004; Muckstadt, 2005) and maintenance optimisation (e.g. Dinesh Kumar et al., 2000), relatively little has been published on the interaction between maintenance, spares and repairs. Most of the current literature deals with the interaction of two out of these three components in a specific setting.

The combination of maintenance and spare parts has been analysed by several authors. For example, Kabir and Al-Olayan (1996), Kabir and Farrash (1996) and Park and Park (1986) deal with an age-based maintenance strategy and non-repairable components. Brezavšček and Hudoklin (2003) present a model with a joint optimisation of a block replacement interval and the maximum inventory level. In Chelbi and Aït-Kadi (2001) the block replacement interval, the optimal stock level as well as the replenishment cycle are optimised simultaneously. Again the components are not repairable, which is encountered in most models that are concerned with joint optimisation of a maintenance policy and a spares provisioning policy.

For the interaction between spare parts and repair capacity, some models have been developed as well. Finite repair capacity is usually modelled by (multi-class) multi-server queues. Gross et al. (1985) were among the first to realise that the combination of inventory and queueing models might lead to insights in the trade-off with respect to maintenance flexibility achieved either through stocks or through sufficient capacity. They attempt to find a cost-optimal combination of the number of spare parts and the number of repair channels, under the constraint that a target service level is met. Kim et al. (2000) have presented an iterative algorithm to determine a cost optimal combination of repair capacities and spare part levels in a single item, multi-echelon model. Avsar and Zijm (2003) consider more general multi-echelon resource structures in which each repair facility may be a queueing network, and show how under Poisson failure rates stock levels at all echelons can be optimised. A similar approach can be used for multi-indenture structures and for combinations of multi-echelon and multi-indenture structures, see Zijm and Avsar (2003). Sleptchenko (2002) deals with the optimisation of the number of spare parts and repair capacity in a multi-item system. Sleptchenko et al. (2005) show that repair priorities may seriously reduce the spare parts investment needed to obtain a target supply availability.

Although the importance of integrating the maintenance strategy, spare parts and repair capacity is recognised, only a few papers describe quantitative models. Natarajan (1968) considers a single unit with spares and a number of repair facilities. By calculating the time to failure the availability is determined. Furthermore, Wang (1993, 1995) consider a single system consisting of operational and stand-by components. They optimise simultaneously the number of stand-by components, number of spares and the number of repairmen. These models are the ones that come the closest to our problem definition. The strongest resemblance is found in Wang (1993) in which there is a number of operating units, a number of warm stand-by units and a number of cold stand-by units (i.e. spare units). Choosing the failure rate of the operating and warm stand-by units to be equal, we have a redundant system in which replacements are done after each component failure (one warm stand-by component turns into an operating unit and a cold stand-by unit becomes warm stand-by). However, they do not cover the interactions we consider in this paper. They do consider a parameter affecting the time until a system failure, namely, the number of warm stand-by units; but they do not have a parameter for the maintenance frequency. So, the number of maintenance set-ups is fixed (maintenance is done after every unit failure). As a consequence the cost involved with the maintenance set-ups is fixed. In this paper we do consider the maintenance frequency as a parameter and we can influence the total maintenance costs by choosing the maintenance frequency.

3. Single system

We use the following model for a single k-out-of-N system. Maintenance is initiated when the system has $m \leqslant N-k+1$ failed components. After a deterministic leadtime L (which can also be equal to zero), maintenance is performed. The decision variables are the maintenance initiation level m, the spare parts stock level S and the repair capacity c. The expected costs per time unit $C_{m,S,c}$ include (i) the holding and depreciation costs of a spare part per time unit C_{spare}, (ii) the cost of repair capacity per time unit C_{capacity}, (iii) the maintenance set-up cost C_{init}. The goal is to minimise these costs $C_{m,S,c}$ given a lower bound Av^* for the expected operational availability $Av_{m,S,c}$. So, we formulate our problem as

$$\begin{aligned} \min \quad & C_{m,S,c} = \frac{C_{\text{init}}}{E[T_m]+L+E[D_{m,S,c}]} + SC_{\text{spare}} + cC_{\text{capacity}} \\ \text{s.t.} \quad & Av_{m,S,c} \geqslant Av^* \end{aligned} \tag{1}$$

Here $E[T_m]$ denotes the expected time until maintenance initiation (at m failed components) and $E[D_{m,S,c}]$ the expected maintenance duration. For the approximation

of the relevant performance measures as a function of the decision variables, we refer to De Smidt-Destombes et al. (2004). There, we have found in numerical experiments that the approximation error in the system availability is relatively small, namely 0.2% on average.

As a first option to solve 1, we consider a straightforward extension of METRIC (Section 3.1), the well-known greedy heuristic for spare part optimisation with maximum increase in availability per invested dollar spare part inventory investment, see Sherbrooke (2004). It turns out that such an approach yields inferior results (Section 3.2). Therefore, we develop a second method where we combine multiple marginal analysis steps in order to find a near-optimal parameter setting (Section 3.3). In Section 4, we compare both methods in a numerical experiment.

3.1. Marginal analysis

Our marginal analysis is a METRIC-like iterative procedure, starting with an initial setting for the decision variables (m, S, c). In each iteration, we consider a marginal change of each decision variable and we select the change leading to the largest quotient of the increase in availability and the cost increase. Because the operational availability is an increasing function of the decision variables S and c, it is logical to select $S=0$ and $c=1$ as initial values. For the maintenance initiation level m, this is not immediately clear.

Given a certain combination (S, c) of the number of spares and repair capacity, we either have a function for which the target availability Av^* cannot be reached (for instance $S=4$, $c=1$ and $S=5$, $c=1$) or a function that has one or multiple points at which the target availability Av^* is reached (the other parameter combinations in Fig. 1). In the first case, we need to increase the number of spares and/or the repair capacity. In the second case, we usually have multiple options for m. Then we should choose for the largest maintenance interval (i.e. largest m) for which the target availability is reached, because then the set-up costs per time unit are lowest. So we start our marginal analysis with the largest realistic value for m and consider the option of decreasing m by one during the greedy algorithm. To compute the largest realistic value for m, we have to take into account possible downtime during the leadtime L. Therefore, an upper bound for m is the value for which we are able to reach the target availability Av^* if the maintenance duration would be zero. This results in the upper bound $m_{\max}$:

$$m_{\max} = \max\left\{1 \leqslant m \leqslant N-k+1 \,\middle|\, \frac{E[T_m]+E[U_m]}{E[T_m]+L} \geqslant Av^*\right\} \quad (2)$$

where U_m denotes the uptime during the leadtime L if maintenance is initiated when m components have failed (see De Smidt-Destombes et al., 2004, for the computation of $E[U_m]$). So we use $S=0$, $c=1$ and $m=m_{\max}$ as initial setting of our decision variables. Next, we use a similar marginal analysis as METRIC, where we consider decreasing m, increasing S and increasing c in each step of the algorithm and select the option yielding the highest increase in availability relative to the additional investment. While performing this algorithm, we may either encounter one or more options for which $Av \geqslant Av^*$ and another option with the largest increment of the availability per cost unit and $Av < Av^*$. Then, we select the latter option and also store the cheapest parameter setting satisfying $Av \geqslant Av^*$. We found that this intermediate solution may be better than the final solution found. So, our first algorithm consists of the following steps:

Step 1: Initialise $S=0$, $c=1$ and $m=m_{\max}$.
Determine $Av_{m,S,c}$ and $C_{m,S,c}$.

Step 2a: Determine $Av_{m-1,S,c}$, $Av_{m,S+1,c}$ and $Av_{m,S,c+1}$.
Determine $C_{m-1,S,c}$, $C_{m,S+1,c}$ and $C_{m,S,c+1}$.

Step 2b: Choose parameter setting $(x,y,z) \in \{(m-1, S, c), (m, S+1, c), (m, S, c+1)\}$ where $(Av_{x,y,z} - Av_{m,S,c})/(C_{x,y,z} - C_{m,S,c})$ is maximal.

Step 3a: If one or more parameter settings yield $Av \geqslant Av^*$, then store the cheapest.

Step 3b: Choose $(m,S,c) = (x,y,z)$, $Av_{m,S,c} = Av_{x,y,z}$ and $C_{m,S,c} = C_{x,y,z}$.
If $Av_{m,S,c} < Av^*$ then go to Step 2a else go to Step 4.

Step 4: Choose the cheapest parameter setting from Steps 2b and 3a.

Unfortunately, some numerical experiments revealed that this algorithm may yield solutions that are far from optimal, see Section 4. A cost difference of 10–20% is not uncommon and in the worst case even a deviation of 171% occurred! In the next section, we analyse the causes of this problem and develop an alternative heuristic that avoids local optima that are much worse than the global optimum.

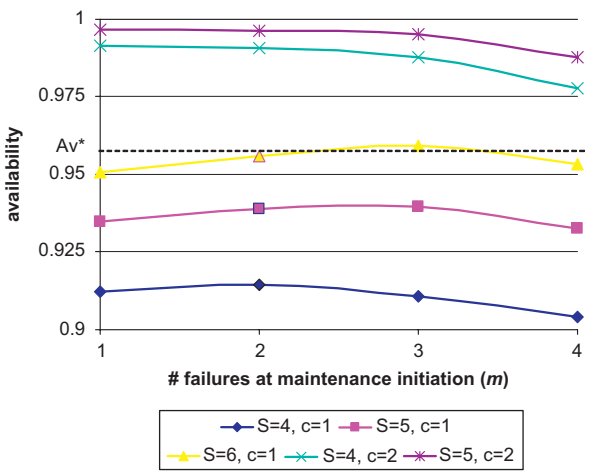

Fig. 1. Example of the availability as function of the number of failures at maintenance initiation (7-out-of-10 system with $L=40$, $\lambda=0.0008$, $\mu=0.001$). The dotted line indicates a possible value for Av^*.

3.2. Drawbacks marginal analysis

The key cause for the bad performance of the marginal analysis is the non-concavity of the expected availability $Av_{m,S,c}$ in the decision variables m, c and S. We distinguish

four major issues that oppose a good performance of the first algorithm:

1. step size of the repair capacity c;
2. choice of the initial parameter setting in the algorithm;
3. shape of the availability as function of the maintenance initiation level m;
4. overestimation of either spares S or repair capacity c;

The *first issue* arises from the large impact of an increase in repair capacity on both costs and availability if the repair shop capacity c is small and the repair shop utilisation is high. For example, suppose that we have a repair shop utilisation of 0.95 when $c=1$. An increase to $c=2$ means a decrease in utilisation to 0.475, which has an enormous impact on the repair shop throughput times. Such an effect can hardly be called "marginal". Besides, it is plausible that an optimal repair shop utilisation may be around 0.6–0.8, which values are not even considered in this example. We can solve this problem by allowing c to have non-integer values (see Sleptchenko, 2002). For practical purposes we can interpret this as e.g. part time work or overtime. Then we can use a step size of (for example) 0.1 full time equivalent (fte) instead of 1 fte. Similarly, we can also choose for only integer values of c and decrease the repair rate with a factor of 10 as well as the cost for capacity, so that the minimum capacity is (for example) $c=10$. In the next section, we discuss how to compute this minimum capacity.

The *second issue*, concerning the initial parameter values, is encountered if the number of spares is small, say far less than the expected number of spares needed for replacement during maintenance. Then the amount of spares is far insufficient and the marginal impact of an extra spare on the availability may be small. Consequently, it is not likely that the algorithm chooses to add a spare. Instead, we see an increase in repair capacity c or a decrease of m. However, when the number of spares would have been larger, the marginal impact on the availability would be higher and so it would be attractive to buy more spares. When we increase the number of spares further, the marginal impact on the availability decreases as expected. Therefore, the availability is not a concave function of the number of spares, as is needed for the marginal analysis.

In order to tackle this issue, Rustenburg (2000) suggests starting values for the number of spares. These starting values are related to the average number of spares in the pipeline at the time of a spare demand. In fact, it means that the starting values are such that the safety stocks are approximately zero. It is plausible that the optimal safety stocks are usually non-negative. In our model, zero safety stock corresponds to a number of spares equal to the expected number of failed components in the system when maintenance starts. However, the corresponding stock level S increases in the maintenance initiation level m (assuming c to be constant). So the initial value of S depends on $m=m_{\max}$, and when m decreases during the execution of the algorithm, the current value of S can be above the initial level for the new value of m. As a consequence, S can have a value above the new initial stock level, even if no spares have been added during the course of the algorithm, and therefore S can be higher than the optimal level, as we encountered in our numerical experiments. Unfortunately, we will never find the optimal value using the marginal analysis, because S can only be *in*creased and cannot be *de*creased. Hence, simply defining initial values for S as the stock levels correspond to zero safety stocks does not solve our problem.

We illustrate the *third issue*, the shape of the availability as function of the maintenance initiation level m, using Fig. 1. Suppose that we found an intermediate solution $S=4$ and $c=1$, where $m=2$ yields the highest availability. When increasing the spares by one ($S=5$, $c=1$) the highest availability is attained for the maintenance initiation level $m=3$. This means less frequent maintenance and therefore less set-up costs. However, the algorithm does not permit an *in*crease of m. As a result, we will not find the optimal parameter setting.

As a *fourth issue*, we found that the algorithm tends to increase the repair capacity in the first iterations when the value of m is still relatively high. This is logical, because a high value of m means infrequent maintenance and hence lumpy demand for repair capacity at the repair shop (infrequent arrival of a large batch of repair jobs). This causes long throughput times, and so the added value of additional capacity is relatively high. However, when the value of m decreases during the execution of the algorithm, the demand for repair capacity becomes more regular and hence *less* repair capacity is needed to attain similar throughput times. So in fact, we should *de*crease the repair capacity, but the marginal approach only allows an *in*crease. As a consequence, we find a repair capacity c that is too high. A similar effect is seen with the number of spares.

We conclude that we can only easily deal with the first issue in the standard marginal approach, but not with the other three issues. Therefore, we have to develop an alternative method.

3.3. Adjusted marginal analysis

To deal with the problems identified in the previous section, we propose the following adjustments:

1. Smaller step sizes for the capacity (first issue).
2. Small initial value for the maintenance initiation level ($m=1$) to enable small initial values for S and c (second issue).
3. Examining high values of m to avoid unnecessary high costs (third issue). Starting with small values of m to solve the second issue concerning the starting values of S and c we will often find a value of m that is smaller than the optimum, see the discussion on the third issue.
4. Balancing the number of spares and repair capacity to reduce costs (fourth issue) to prevent ending up with a solution in which the number of spares and/or capacity is higher than necessary.

We developed a new algorithm using these four adjustments. In the remainder of this section we describe the steps of this adjusted marginal analysis algorithm.

3.3.1. Step 0: initialisation

As stated in Section 3.2 we use smaller step sizes for the repair capacity in such a way that we know for sure that $c=1$ implies insufficient capacity. That is, we use a step size of 1 in the optimisation algorithm, which may correspond to (for example) 0.1 fte. As initial value for c, we choose the minimum capacity needed to repair all failed components in the long run at an availability close to the target. The number of component failures per cycle equals m plus the number of component failures during the leadtime $(N-m)(1-e^{-\lambda L})$. Ignoring downtime during the leadtime, we find that we may at most use a period with length $(E[T_m]+L)/Av^*$ to restore the components at rate $c\mu$. Therefore, we find

$$c_{\min}(m)=\left\lceil \frac{m+(N-m)(1-e^{-\lambda L})}{(E[T_m]+L)\mu} Av^* \right\rceil \quad (3)$$

where $\lceil X \rceil$ denotes the smallest integer larger than or equal to X. Unfortunately, this initial value depends on m. For simplicity we use the minimum over all m as initial value for c, so that $c_{\min}=c_{\min}(1)$. To avoid the problems with the first part of the function in S, we choose the expected number of failed components when the system comes in for maintenance as the initial value of S:

$$S_{\min}(m)=\lfloor m+(N-m)(1-e^{-\lambda L}) \rfloor \quad (4)$$

where $\lfloor X \rfloor$ denotes the largest integer smaller than or equal to X. In practice, the target availability is not very low (say 80–90% or even higher), and therefore it is not expected that this initial number of spares is too high. However, this initial value depends on m again. We solve this by choosing an initial value $S=S_{\min}(1)$ that corresponds to maintenance initiation level $m=1$. In this way, we avoid an overestimation of the number spares needed in the optimum. If we increase m during the algorithm, we evaluate whether we violate the lower bound $S_{\min}(m)$ and if so, we increase S simultaneously.

Putting this together, we use as initial values $m=1$, $S=S_{\min}(1)$ and $c=c_{\min}(1)$.

3.3.2. Step 1: improving availability without increasing costs

Here we only consider an increase in m as long as the costs $C_{m,S,c}$ decrease and the availability $Av_{m,S,c}$ increases. The lower bound $S_{\min}(m)$ increases simultaneously with m. To avoid too high values of the capacity in the beginning of the algorithm, the value of $c=c_{\min}(1)$ remains unchanged. In this step, we reduce the maintenance set-up costs (decreasing maintenance frequency) but we increase the spare part inventory costs. As we see from Fig. 1, the combination of increasing m and S initially leads to an increase in availability. Therefore, we proceed as long as the net effect is a cost reduction and an increase in availability. So, in the first part of this step (step 1a) we determine the availability and costs corresponding to an increase of m (and possibly an increase of S as well). The second part of this step (step 1b) consists of adjusting the parameters as long as the availability increases and the costs decrease. The resulting values for m, S and c are starting values for the next step.

3.3.3. Step 2: improving availability until Av^* with acceptance of increasing costs

If we have already reached the target availability Av^*, we move to step 3. Otherwise, we apply a marginal analysis approach in which we consider an increase of the repair capacity and an increase of spares. In step 2a we consider the following two options:

- increase c by one and simultaneously increase the value of m as much as possible such that the availability does not decrease compared to the availability we found thus far. Note that we modify (increase) the number of spare parts S if the increase in m causes a violation of the spare part lower bound $S_{\min}(m)$. As an example of this option from Fig. 1, consider the parameter setting $m=2$, $c=5$ and $S=1$. If we increase the capacity to $c=6$, we could increase m to $m=4$ instead of $m=2$, thereby reducing costs without loss of availability;
- increase S by one and simultaneously increase m as much as possible such that the availability increases compared to the availability we found thus far.

In step 2b we choose one of these options as the new parameter setting. Both options may cause an increase of the costs as well as a decrease of the costs. In case of a cost reduction ($\Delta C<0$) we choose the option with the lowest, most negative, value for $\Delta Av/\Delta C$. Otherwise ($\Delta C>0$) we choose the option with the largest $\Delta Av/\Delta C$.

We repeat this step until we reach or exceed the target availability level Av^*.

3.3.4. Step 3: reducing costs by increasing m and maintaining Av^*

Now we have reached the target availability, but probably not at minimum costs. Therefore, we look for other solutions having a similar availability but lower costs by increasing the maintenance initiation level m. Without this step we often end up with a value of m that is too small, because we started our algorithm with $m=1$ (see Fig. 1). Basically, we continue the previous step, but now we accept cost reductions only. Also, we accept all availability levels $\geqslant Av^*$. This adjustment solves the problems mentioned under issue two.

3.3.5. Step 4: balancing the parameter setting

Finally, we address the third issue about possible compensation between the spare part inventory level S and the repair capacity c. We perform a last step to find a better balance between the parameter values, where we also include the value of m. We consider four options to reduce the costs while attaining the target availability level. Each option consists of a modification in two parameters simultaneously, where one parameter modification yields a cost increase and the other yields a cost

decrease. As long as the overall cost impact is a decrease, we improve our solution.

- The first option is to decrease the capacity by one (decrease in repair capacity costs) and increase the number of spares (increase in spare part inventory costs), where we choose a minimal increase in S such that the availability is at least equal to Av^*.
- The second option is to decrease the capacity by one (decrease in repair capacity costs) and decrease the value of m as much as necessary in order to obtain Av^* (increase in set-up costs).
- The third and fourth options are analogous to these two options, only then the number of spares is decreased by one, with a necessary increase of the capacity or decrease of the maintenance initiation level.

After determining the parameter settings for each option in step 4a, we choose from these options the one that has the largest cost reduction in step 4b. We repeat this procedure as long as we can find a cost reduction.

Summarised, our enhanced algorithm consists of the following steps:

Step 0: Initialise $m=1$, $S=S_{\min}(1)$ (Eq. (4)) and $c=c_{\min}(1)$ (Eq. (3)).
Determine $Av_{m,S,c}$ and $C_{m,S,c}$.

Step 1a: Determine $Av_{m+1,S_{\min}(m+1),c}$ and $S_{m+1,S_{\min}(m+1),c}$.

Step 1b: If $Av_{m+1,S_{\min}(m+1),c} > Av_{m,S_{\min}(m),c} \wedge C_{m+1,S_{\min}(m+1),c} < C_{m,S_{\min}(m),c} \wedge m+1 < m_{\max}$
then $(m,S,c)=(m+1,S_{\min}(m+1),c)$ and go to Step 1a else go to Step 2a.

Step 2a: If $Av_{m,S,c} \geqslant Av^*$ go to Step 3 else
Find $\max \widetilde{m}_S \in [m, m_{\max}]$ with $Av_{m,S,c} < Av_{\widetilde{m}_S,S+1,c}$.
Find $\max \widetilde{m}_c \in [m, m_{\max}]$, $\widetilde{S}_c = \max\{S, S_{\min}(\widetilde{m}_c)\}$ with $Av_{m,S,c} < Av_{\widetilde{m}_c,\widetilde{S}_c,c+1}$.

Step 2b: If $\min\{C_{\widetilde{m}_S,S+1,c}, C_{\widetilde{m}_c,\widetilde{S}_c,c+1}\} < C_{m,S,c}$
choose $(x,y,z) \in \{(\widetilde{m}_S, S+1, c), (\widetilde{m}_c, \widetilde{S}_c, c+1)\}$ with $\min(Av_{x,y,z} - Av_{m,S,c})/(C_{x,y,z} - C_{m,S,c})$.
Else $(x,y,z) \in \{(\widetilde{m}_S, S+1, c), (\widetilde{m}_c, \widetilde{S}_c, c+1)\}$ with $\max(Av_{x,y,z} - Av_{m,S,c})/(C_{x,y,z} - C_{m,S,c})$.
Go to Step 2a.

Step 3: Find $\max \widetilde{m}_S \in [m, m_{\max}]$ with $Av_{\widetilde{m}_S,S+1,c} \geqslant Av^*$.
Find $\max \widetilde{m}_c \in [m, m_{\max}]$, $\widetilde{S}_c = \max\{S, S_{\min}(\widetilde{m}_c)\}$ with $Av_{\widetilde{m}_c,\widetilde{S}_c,c+1} \geqslant Av^*$.
If $\min\{C_{\widetilde{m}_S,S+1,c}, C_{\widetilde{m}_c,\widetilde{S}_c,c+1}\} < C_{m,S,c}$ choose cheapest and go to Step 3.
Else go to Step 4a.

Step 4a: Determine S_c with $Av_{m,S_c,c-1} \geqslant Av^*$ and $Av_{m,S_c-1,c-1} < Av^*$.
Determine m_c with $Av_{m_c,S,c-1} \geqslant Av^*$ and $Av_{m_c+1,S_c,c-1} < Av^*$.
Determine c_S with $Av_{m,S-1,c_S} \geqslant Av^*$ and $Av_{m,S-1,c_S-1} < Av^*$.
Determine m_S with $Av_{m_S,S-1,c} \geqslant Av^*$ and $Av_{m_S+1,S-1,c} < Av^*$.

Step 4b: If $\min\{C_{m,S_c,c-1}, C_{m_c,S_c,c-1}, C_{m,S-1,c_S}, C_{m_S,S-1,c}\} < C_{m,S,c}$
then $C_{m,S,c} = \min\{C_{m,S_c,c-1}, C_{m_c,S_c,c-1}, C_{m,S-1,c_S}, C_{m_S,S-1,c}\}$ and go to Step 4a

Compared to the simple marginal analysis algorithm from Section 3.1, the number of availability computations has increased. However, we usually find a solution that is much closer to the optimum. Next, we discuss the quality of this method and its computational performance.

4. Numerical results for the single system

We study three system sizes: 7-out-of-10 systems, 58-out-of-64 systems and 2700-out-of-3000 systems. For each system, we consider 108 parameter combinations for repair times and cost parameters. We consider the parameters that we initially used for the marginal analysis algorithm from Section 3.1. For the adjusted algorithm (Section 3.3), we divided the repair rate as well as the cost for capacity by 10. In this way, we start at a higher value for $c_{\min}(1)$ and therefore the relative step size for the repair capacity is smaller. In our comparison between both algorithms, we only use the adjusted input parameters, which are given in Table 1. The cost parameters are given per time unit. For the failure rate we choose $\lambda = 0.0001$ for all systems. The leadtime equals $L=168$ for the 2700-out-of-3000 system and $L=40$ for the other two systems. We use a target availability of $Av^* = 0.99$.

We used (time consuming) enumeration as benchmark. To this end, we need upper and lower bounds for each of the three parameters. For m, we obviously search over $m \in [1, m_{\max}]$. Lower bounds for S and c are $S_{\min}(1)$ and $c_{\min}(1)$, respectively. However, it is not immediately clear how to choose the corresponding upper bounds. Therefore, we proceed as follows. First, we look for an arbitrary parameter setting that satisfies the availability restriction Av^*. We chose $m = \max\{1, \lfloor 0.5 m_{\max} \rfloor\}$ and $c = c_{\min}(m)$ and find the minimum number of spares S needed to obtain Av^*. Next, we use the corresponding cost $\widehat{C}_{m,S,c}$ to find upper bounds for S and c. As the total costs of spares in the optimum solution should be less than $\widehat{C}_{m,S,c}$, upper bounds for S and c are given by $\widehat{C}_{m,S,c}/C_{\text{spare}}$ and $\widehat{C}_{m,S,c}/C_{\text{cap}}$, respectively. To reduce the computational effort of enumeration, we recalculate these upper bounds each time we find a better solution during enumeration.

For each system size, we show in Table 2 the mean and maximum relative deviation from the optimal costs per time unit $C^*_{m,S,c}$. Besides, we show the percentage of scenarios in which the optimisation heuristic found exactly the optimal solution.

Looking to our detailed results (not included in this paper because of the amount of data), we observe that increasing the cost for capacity results in a decrease of capacity compensated by more spares and sometimes combined with a shift in the maintenance frequency. If the cost for spares increases we see that the first result is a lower maintenance initiation level often combined with an increase of the repair capacity. An increase of the maintenance initiation costs is compensated by an increase of the maintenance initiation level combined with an increase of the spares amount. The repair capacity remains unchanged in almost every scenario. For all

Table 1
For different system sizes we used different input parameters, resulting in 108 scenarios per system size

	μ	C_{init}	C_{spare}	C_{cap}
7-out-of-10	0.00005, 0.000075, 0.0001	50 000, 75 000, 100 000	0.5, 1, 2.5, 5	10, 15, 30
58-out-of-64	0.0005, 0.00075, 0.001	50 000, 75 000, 100 000	0.5, 1, 2.5, 5	10, 15, 30
2700-out-of-3000	0.003, 0.015, 0.03	50 000, 75 000, 100 000	0.5, 1, 2.5, 5	10, 15, 30

Table 2
For different system sizes the mean and maximum cost differences are given for the simple and adjusted marginal analysis algorithms compared to enumeration

	Simple marginal analysis			Adjusted marginal analysis		
	Mean diff. (%)	Max. diff. (%)	Opt. found (%)	Mean diff. (%)	Max. diff. (%)	Opt. found (%)
7-out-of-10	6.07	13.1	21.3	0.10	2.5	91.7
58-out-of-64	13.32	43.4	13.0	0.15	1.5	75.9
2700-out-of-3000	29.13	171.0	0.0	0.18	3.2	32.4

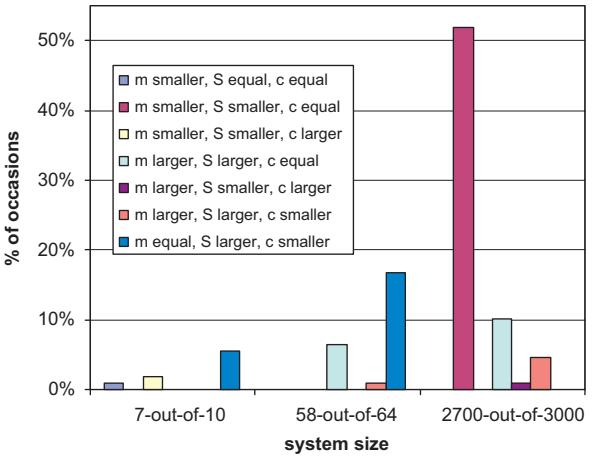

Fig. 2. Deviations from the optimal solution found with the adjusted marginal analysis for different system sizes. The percentages given are the percentages from the total number of scenarios, so including the scenarios in which the optimal solution was found.

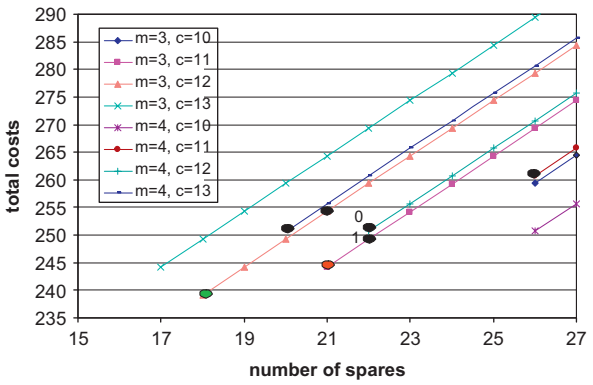

Fig. 3. A schematic representation of the balancing step in the optimisation heuristic for a 7-out-of-10 system with $L = 40$, $\lambda = 0.0001$, $\mu = 0.0001$, $C_{init} = 100\,000$, $C_{spare} = 5$ and $C_{cap} = 10$. All parameter settings that are given satisfy the target availability level. The optimal solution is represented by the green dot, while the red dot is the sub-optimal solution we find.

scenarios we see, independent of the cost parameters, that the maintenance initiation level is such that the system does not fail before arriving at the repair shop. So, the maintenance policy is obviously to perform preventive maintenance.

As can be seen in Table 2, the enhanced algorithm yields much better solutions than the straightforward marginal approach. Using our enhanced algorithm we also find the exact optimum solution more frequently. For the cases in which the parameter setting of the adjusted marginal analysis differs from the optimal solution, we can classify the type of deviation, see Fig. 2. It shows the percentage of each type of deviation as a percentage of the total number of scenarios. We see that for the large systems we find too small values for m and S in most cases. For smaller systems, we tend to find the optimal value of m combined with too large values for S and too small values for c.

The deviations that are relatively large, more than 1%, are mainly caused by too many spares and too few capacity. In these cases we end up in a local minimum from which we do not reach the global minimum by balancing the spares and capacity using our heuristic. In Fig. 3 we illustrate the balancing step of the algorithm. The parameter setting marked as 0 is the solution we find after the third step. From there we have the possibility to move to one of the black dots if this improves the cost. Obviously, the option marked as 1 is the cheapest. Finally, we end up in the red dot instead of the green one. Unless we would except more expensive solutions, we are not able to make the change to the line with $m = 3$, $c = 12$.

When we consider the utilisation rates, the differences between the optimal solution and the solution found with our optimisation heuristic are small as can be seen from Table 3. The deviating utilisation rates are always found in the scenarios where the parameter setting of the adjusted marginal analysis has a too large number of spares and a too small number of capacity (sometimes combined with a maintenance initiation level that is too large). So, we

Table 3
For different system sizes the average utilisation rates are given for the solutions found using enumeration and the adjusted marginal analysis

	Enumeration			Adjusted marg. analysis		
	Min. (%)	Mean (%)	Max. (%)	Min. (%)	Mean (%)	Max. (%)
7-out-of-10	73.5	81.4	83.3	78.4	81.8	89.0
58-out-of-64	61.0	85.3	93.8	74.0	86.9	93.8
2700-out-of-3000	81.5	92.4	97.6	81.5	92.8	98.0

Table 4
For different system sizes the average number of availability computations is given for the enumeration, the simple marginal analysis and the adjusted marginal analysis algorithm

	Enumeration	Marg. analysis	Adjusted marg. analysis
7-out-of-10	442	85	87
58-out-of-64	2348	45	73
2700-out-of-3000	658 361	826	1249

may conclude from this table that the utilisation rate is not affected very much if we do not find the optimal parameter setting in all cases.

In Table 4 we show the average number of availability computations per solution method. We see that the additional computational effort for the optimisation algorithm remains within reasonable bounds. Although enumeration is an option for small systems, it becomes cumbersome for large systems. Especially, since the computation times (on a Pentium III 996 MHz) for the large systems become almost 7.5 h for 108 scenarios (compare to 3.8 min using the optimisation algorithm). Of course, one can argue that it is possible for large systems to do a rougher enumeration (say a step size 5) for the parameters S and m and then do a more extensive enumeration for a few of the best solutions. However, for this heuristic to be quicker than the one we propose, the number of computations needs a reduction of more than 99.8% of the enumeration we performed.

There is, however, one disadvantage when using this optimisation algorithm. The algorithm finds a near-optimal solution, but not via a path of near-optimal solutions for various target availability levels as is true for METRIC. This property of METRIC can be used to construct an availability–cost trade-off curve. As a consequence, in principle we have to start our computations all over again if the target availability level changes. Of course, one could use the solution found for a certain target availability as initial value to find the best solution for a somewhat higher target availability, just like METRIC. However, some experiments revealed that this may lead to inferior results.

5. Conclusions and further research

In this paper, we presented a heuristic method to find a cost effective balance between maintenance frequencies, spare parts inventories and repair capacity in order to achieve a target availability level. We considered a single k-out-of-N system under condition-based maintenance. We showed that "simply" extending the METRIC approach yields inferior results, since the relationship between the decision parameters and the operational availability is not a monotonous one. We identified four major issues why this approach does not work and found a solution to deal with those issues.

We compared results of our optimisation heuristic to the results of a complete enumeration. We found that the cost differences are limited to 0.2% on average for the single k-out-of-N system.

It is relatively easy to modify our optimisation algorithm for several model variants, such as the inclusion of component wear-out and the extension to an installed base of k-out-of-N systems under block replacement. For the latter model, we found that our optimisation heuristic yields costs that are on average 0.8% more than the costs found by enumeration. At the same time, the computation times were a lot smaller, minutes compared to hours or even days. We refer to De Smidt-Destombes (2006) for the details.

Relevant further research should cover a generalisation to multi-item models which is certainly not straightforward. Also, the inclusion of a multi-indenture product structure (e.g. multiple k-out-of-N systems within one system) is relevant but not simple. A simple but also relevant generalisation is alternative stand-by models (warm and cold stand-by instead of hot stand-by). For a discussion on the possibilities and issues regarding these research directions, we refer to De Smidt-Destombes (2006).

References

Avsar, Z., Zijm, W., 2003. Capacitated Two-Echelon Inventory Models for Repairable Item Systems, vol. 60. Kluwer Academic Publishers, Dordrecht.

Brezavšček, A., Hudoklin, A., 2003. Joint optimization of block-replacement and periodic-review spare-provisioning policy. IEEE Transactions on Reliability 52 (1), 112–117.

Chelbi, A., Aït-Kadi, D., 2001. Spare provisioning strategy for preventively replaced systems subjected to random failure. International Journal of Production Economics 74, 183–189.

De Smidt-Destombes, K., 2006. Spares and repairs for maintaining redundant systems. Ph.D. Thesis.

De Smidt-Destombes, K., Van Der Heijden, M., Van Harten, A., 2004. On the availability of a k-out-of-N system given limited spares and repair capacity under a condition based maintenance strategy. Reliability Engineering and System Safety 83 (3), 287–300.

De Smidt-Destombes, K., Van der Heijden, M., Van Harten, A., 2006. On the interaction between maintenance, spare part inventories and repair capacity for a k-out-of-n system with wear-out. European Journal of Operational Research 174 (1), 182–200.

De Smidt-Destombes, K., Van der Heijden, M., Van Harten, A., 2007. Spare parts analysis for k-out-of-n systems under block replacement and finite repair capacity. International Journal of Production Economics 107 (2), 404–421.

Dinesh Kumar, U., Crocker, J., Knezevic, J., El-Haram, M., 2000. Reliability, Maintenance and Logistic Support: A Life Cycle Approach. Kluwer Academic Publisher, Dordrecht.

Gross, D., Miller, D., Soland, R., 1985. On common interests among reliability, inventory and queuing. IEEE Transactions on Reliability R-34 (3), 204–208.

Kabir, A., Al-Olayan, A., 1996. A stocking policy for spare part provisioning under age based preventive replacement. European Journal of Operational Research 90, 171–181.

Kabir, A., Farrash, S., 1996. Simulation of an integrated age replacement and spare provisioning policy using SLAM. Reliability Engineering and System Safety 52 (2), 129–138.

Kim, J., Shin, K., Park, S., 2000. An optimal algorithm for repairable-item inventory system with depot spares. Journal of Operations Research Society 51, 350–357.

Muckstadt, J., 2005. Analysis and Algorithms for Service Parts Supply Chains. Springer, Berlin ISBN: 0-387-22715-6.

Natarajan, R., 1968. A reliability problem with spares and multiple repair facilities. Operations Research 16 (5), 1041–1057.

Park, Y., Park, S., 1986. Generalized spare ordering policies with random lead time. European Journal of Operational Research 23, 320–330.

Rustenburg, W., 2000. A system approach to budget-constrained spare parts management. Ph.D. Thesis, BETA Research Institute.

Sherbrooke, C., 2004. Optimal Inventory Modeling of Systems: Multi Echelon Techniques, second ed. Kluwer Academic Publishers, Dordrecht ISBN: 1-402-07849-8.

Sleptchenko, A., 2002. Integral inventory control in spare parts networks with capacity restrictions. Ph.D. Thesis, BETA Research Institute. ISBN: 90-365-1817-2.

Sleptchenko, A., Van Der Heijden, M., Van Harten, A., 2002. Effects of finite repair capacity in multi-echelon, multi-indenture service part supply systems. International Journal of Production Economics 79, 209–230.

Sleptchenko, A., Van Der Heijden, M., Van Harten, A., 2003. Trade-off between inventory and repair capacity in spare part networks. Journal of the Operational Research Society 54 (3), 263–272.

Sleptchenko, A., Van der Heijden, M., Van Harten, A., 2005. Using repair priorities to reduce stock investment in spare part networks. European Journal of Operational Research 163 (3), 733–750.

Wang, K., 1993. Cost analysis of the M/M/R machine-repair problem with mixed standby spares. Microelectronics and Reliability 33 (9), 1293–1301.

Wang, K., 1995. An approach to cost analysis of the machine repair problem with two types of spares and service rates. Microelectronics and Reliability 35 (11), 1433–1436.

Zijm, W., Avsar, Z., 2003. Capacitated two-indenture models for repairable item systems. International Journal of Production Economics 81–82 (C), 573–588.