Journal of Statistical Planning and Inference 142 (2012) 2409–2420

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Optimal design and maintenance of a repairable multi-state system with standby components

Ramin Moghaddass, Ming J. Zuo*, Mayank Pandey

Reliability Research Lab, Department of Mechanical Engineering, University of Alberta, Edmonton, Canada

ARTICLE INFO

Article history:
Received 11 September 2011
Received in revised form
14 February 2012
Accepted 18 February 2012
Available online 24 February 2012

Keywords:
Optimal design
Maintenance strategies
Repairable multi-state system
Redundancy

ABSTRACT

The configuration of a repairable system directly influences its performance measures, such as mean time between failures and steady state availability. Additionally, maintenance strategies such as corrective, preventive, and condition-based can affect the performance of the system. The objective of this work is to investigate the trade-offs between the configuration of a repairable multi-state system with binary components and its maintenance strategy. The corresponding stochastic process for the proposed model is formulated using the continuous-time Markov process and important performance measures of such a multi-state system are derived. An optimization model is introduced for the cost-effective design of this repairable multi-state system. The results are demonstrated using a numerical example for a power generation system.

1. Introduction

In today's technological environment, it is essential to conduct reliability analysis for complex systems. When systems are repairable, the main measure of interest is availability which needs to be maximized to achieve the optimal performance (Lad et al., 2008). The most common approach to improve the availability and guarantee the continuous operation of the system is to employ redundancy, either in the form of active redundancy or standby redundancy. One of the most common forms of redundancy is the k-out-of-n:G structure which has a wide range of applications (Moghaddass et al., 2011a). Under active redundancy, redundant components operate the same way as other active components. Standby redundancy is further classified as hot standby, warm standby, and cold standby redundancy (Amari and Dill, 2009).

In hot standby redundancy, all non-failed components are subject to failure unless they need to be suspended when the system is down. In contrast, cold standby or alternatively called spares may fail only when switched into the active position. Warm standby components are subject to failure with a failure rate less than the failure rate of active components. We note here that when there are no switching delays and failures, hot standby redundancy and active redundancy are equivalent (Amari and Dill, 2009). In this paper, we consider active redundancy and cold standby redundancy only; however our results for active redundancy remain the same for hot standby components with no switching delays and failures.

Many of the reported work on redundancy considered systems with only two states. In many practical cases, to satisfy different levels of demand, a system may work in different efficiency levels, resulting in a multi-state system (MSS) (Lisnianski and Levitin, 2003). Among different types of MSS, MSSs with binary components have received a significant amount of

* Corresponding author.
E-mail address: Ming.Zuo@UAlberta.ca (M.J. Zuo).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2012.02.045

attention due to their applications. Examples of such systems can be found in Levitin and Lisnianski (2000), Liu and Huang (2010), and Machani and Nourelafath (2010). The MSS considered in this paper is similar to the one in Huang et al. (2000), wherein maintaining at least a certain system-state level might require a different number of components to be at a certain state or above. For a repairable MSS, using active and hot standby redundancy may result in satisfying more than minimum required demand. However, these redundant components are subject to additional operational and maintenance costs. On the other hand, cold standby components do not deteriorate while in the standby mode. However, the availability of the system may decrease due to the time needed to switch a cold standby component to the active mode.

In addition to redundancy, the availability of a repairable system can also be improved by increasing the maintenance capacity of the system. Although theoretically, unlimited repair capacity is desirable, it costs too much to be practical in most applications. Maintenance frequency also affects the system availability. For the highest availability, the maintenance team should be sent for repair as soon a component fails. This course of action may increase the cost of maintenance setup. On the other hand, if a repair team is sent only after the system has failed, the unavailability can result in a high cost of downtime (Moghaddass and Zuo, 2011). Because of the maintenance initiation setup, it is important to determine the circumstances under which the maintenance setup should be initiated, i.e. it is essential to investigate how and when the components in a repairable system should be replaced or repaired (Chiang and Yuan, 2001).

Many researchers have investigated the optimal system design; however few have considered the optimal design of a repairable system. A comprehensive review on the optimal design for repairable systems can be found in (Lad et al., 2008; Mohamed, 1992). Most have investigated the optimal design with regards to availability, while considering constraints on budget and space. Further, few publications are available for the optimal design of a repairable system with respect to the overall system profit. Pham (1992) considered a simple cost-effective design for a k-out-of-n:G system. Sasaki et al. (1977) investigated the number of cold standby components for a series system with 1-out-of-n:G subsystems. Srivastava and Fahim (1988) introduced a cost-based model to find the number of cold standby components for a series system with k-out-of-n:G subsystems. Fawzi and Hawkes (1991) considered a k-out-of-n:G system with cold standby components, assuming that components are sent for repair immediately upon failure.

There are also some published works on the trade-off between design and maintenance factors. Misra (1974) investigated the trade-off between the number of standby components and the repair capacity for a series system with k-out-of-n:G subsystems. Sharma and Misra (1988) used an intelligent search algorithm to find the optimum values for the number of active redundant components, cold standby components, and repair facilities. Gurov et al. (1995) developed an optimization model to find the number of redundant or cold standby components and repairmen. Amari and Pham (2007) developed a cost-based model to find the number of cold standby components for a repairable system. The trade-off between cold standby components and repair capacity was also analyzed in Sleptchenko et al. (2003).

In a series of work by Krishnamoorthy and Ushakumari (2001), Krishnamoorthy et al. (2002) and Ushakumari and Krishnamoorthy (2004), the repair activation point for a k-out-of-n:G system with standby components (cold, warm, and hot) and a single repair facility was investigated. De Smidt-Destombes et al. (2009) investigated the trade-off between the cold standby components, repair capacity, and maintenance frequency of a repairable k-out-of-n:G system. Wang (1993, 1995) developed an optimization model to find the numbers of cold and warm standby components, and repairmen for a repairable system. De Smidt-Destombes et al. (2006) considered a k-out-of-n:G system, where repair activation point, the cold standby stock level, the repair capacity, and the repair priority setting were considered as the variables to control the availability. De Smidt-Destombes et al. (2004) investigated the trade-off between the availability, cold standby stock level, the maintenance policy, and the repair capacity for a k-out-of-n:G system.

To the best of our knowledge, analyzing the trade-off between maintenance and design factors to simultaneously find the numbers of active redundant components, cold standby components, and repairmen and the maintenance initiation point for a repairable MSS has not been studied in the literature. The objective of this paper is to provide a comprehensive modeling for the cost-effective design of a repairable MSS with binary-state components considering several types of operational and maintenance costs. An advantage of the proposed method over similar work is that we considered a more practical case, in which the switching time for a cold standby component and the maintenance setup time are both random variables. A method will be introduced to generate the system transition matrix and the formula for important performance measures for such a system will be developed. We note here that finding the optimal solution of the proposed optimization model is out of the scope of this paper. Our contribution in this paper is on the modeling of such MSS, in the sense that the trade-offs between design and maintenance factors are considered. This work is an extension of our previous work (Moghaddass and Zuo, 2011), where the system was a binary k-out-of-n:G system and the switching time for cold standby components and maintenance initiation setup time were assumed to be negligible.

The remainder of this paper is organized as follows. Section 2 presents the formulation of the stochastic model including state space, transition diagram, and transition rates. Several performance measures of such a system are introduced in Section 3. The optimization model is presented in Section 4. In Section 5, a numerical example for a power generation system is analyzed. Finally, conclusions and suggestions for future work are presented in Section 6.

2. Model formulation

In this section, the corresponding stochastic model for the multi-state system is described and a systematic approach is presented to generate the transition matrix of the system.

2.1. Assumptions and notation

The notations used in this paper are as follows:

n	number of active redundant components
s	number of cold standby components
r	number of repairmen (repair facilities)
p	critical point of maintenance initiation
M	the number of health states for the system
S_w	wth health state ($1 \le w \le M$) of the system (S_1: highest level of performance)
O_w	minimum number of active components that is required to be healthy in order for the multi-state system to work in the healthy state w ($O_1 > O_2 > \ldots > O_{M-1}$)
E	the performance level (output) of each component when it is in perfect functioning state
d_w	demand level which needs to be satisfied for the wth health state ($d_w = EO_w$)
m	number of suspended active components when the system is down
(i,j,z)	detailed system state; where i is the number of failed active components, j is the number of failed cold standby components, and z is a variable indicating the condition of the system with respect to maintenance ($z=0$ means that the system is not under maintenance, $z=1$ means that the system is waiting for maintenance, and $z=2$ means that the system is under maintenance)
$G(t)$	the performance level (output) of the system
D	the set of all reachable states in the system
$\overline{D}$	the set of all reachable states in which the system is down (complete failure)
N	total number of reachable states in the multi-state system
$\pi_{(i,j,z)}$	steady state probability for state (i,j,z) in the multi-state system
π'_w	steady state probability for the wth health state of the multi-state system
λ	failure rate of the identical components (active components)
μ	repair rate of the identical components (failed components)
δ	switching rate for the cold standby components
ε	maintenance initiation setup rate
c_1	operational cost per unit time for each active component
$c_2^{(1)}$	cost per unit time for each cold standby component (non-functionality cost including holding, depreciation, etc.)
$c_2^{(2)}$	cost per unit time for non-functionality of active component (all costs associated with active components excluding the operational cost c_1)
c_3	fixed cost of each maintenance initiation (fixed cost through maintenance setup)
c_4	cost of hiring each repair facility (repairman) per unit time
c_5	cost of repair per unit time
c_6	cost of purchase for each active and standby components
c'_w	profit per unit time when the system is in health condition S_w
b_1	capital budget available for purchasing active and cold standby components
b_2	available space (maximum number) for active components
b_3	available space (maximum number) for cold standby components

The main assumptions regarding the model are as follows:

(I) Both active redundancy and standby redundancy are considered in the sense that there are n identical active components and s additional identical cold standby components in the system. Active components are in use (operational mode) and are subject to failure, while cold standby components do not age while in the spare mode (cold standby mode). However, a cold standby component can be switched to the operation mode, so that it becomes an active component. These two variables (n and s) are our design decision variables.

(II) All active and standby components are stochastically independent and identical with two states following exponential failure time and repair time distributions.

(III) After an active component fails, a cold standby component (if available) automatically replaces that failed component. Therefore, the cold standby components become an active component and vice versa. The switching time from a cold standby position to an active position follows an exponential distribution with a constant switching rate.

(IV) The system is repairable and functions in a finite number of distinct levels of performance (outputs). Each level of performance is referred to as a health state.

(V) The performance level of the multi-state system at time t ($G(t)$) is defined as $G(t)=\phi(G_1(t),\ldots,G_n(t))$, where $G_i(t)$ is the random variable representing the performance level of the ith active component at time t.

(VI) The maintenance policy is based on the condition of the system in such a way that the maintenance setup is initiated whenever the number of failed active components reaches a critical value p. The repair team including r repairmen

starts repairing all failed cold standby components only. In our work, p and r are maintenance decision variables. The repair team leaves when there is no failed component in the system. Repair cannot be done on a failed active component unless it is switched to a cold standby configuration. The maintenance initiation setup time follows an exponential distribution with maintenance initiation setup rate ε.

(VII) A repaired component becomes as good as new after maintenance (perfect repair).

(VIII) When the system is down (complete failure), m of all non-failed active components are suspended just after the system failure. This assumption covers the shut-off rules of suspended animation, continuous operation, and a combination of these two (Moghaddass and Zuo, 2011; Moghaddass et al., 2011a, 2011b).

2.2. State representation

To demonstrate the detailed system state at each point of time, a triplet (i,j,z) as defined in Section 2.1 is used. With this triplet, the state of the system and its components can be identified at any point of time. Also, by applying certain rules (to be described later in Section 2.3), the transition rates between states can be defined automatically. To simplify analysis of the model based on our assumptions, we need to systematically define all reachable states in the system. We have divided the states of the system into different classes with similar properties. Based on the partial transition diagrams of the system shown in Figs. 1–3, we assign all states in the same row to a set of states called "*class*". These classes of states are distinguished based on two factors: the number of failed cold standby components and the condition of the system with respect to maintenance. These two factors are the same in all states within a same class. To distinguish different classes, we define notation $D_{j,z}$, where j is the number of failed cold standby components, and z is the condition of the system with respect to maintenance.

Classes in the form of $D_{j,0}$, $1 \le j \le s$, contain states with j cold standby components where the system is not under maintenance. Classes in the form of $D_{j,1}$, $1 \le j \le s$, contain states with j cold standby components where the system is not under maintenance and the maintenance setup has already been initiated. Classes in the form of $D_{j,2}$, $1 \le j \le s$, contain states with j cold standby components where the system is under maintenance. The existing classes in each form of $D_{j,0}$, $D_{j,1}$, and $D_{j,2}$ for the introduced repairable multi-state system can be constructed as

$$D_{j,0} = \{(i,j,0)\}, \quad \begin{cases} j=0, \quad i=0 \quad p=1 \\ 0 \le j \le s, \quad 0 \le i \le p-1 \quad p \ne 1 \end{cases}, \tag{1}$$

$$D_{j,1} = \{(i,j,1)\}, \quad \begin{cases} p-j \le i \le n-m \quad 0 \le j \le p \\ 0 \le i \le n-m \quad p \le j \le s \end{cases}, \tag{2}$$

$$D_{j,2} = \{(i,j,2)\}, \quad 0 \le j \le s, \quad 0 \le i \le n-m, \quad i+j \ne 0. \tag{3}$$

All reachable classes of states can be constructed using Eqs. (1)–(3). Also, the three forms of classes shown in Eqs. (1)–(3) cover all reachable states in the system. Therefore, the set of reachable states can be expressed as $D = \bigcup_{0 \le j \le s} D_{j,0} \cup D_{j,1} \cup D_{j,2}$. The total number of states in the system (N) can be calculated under the next three possible

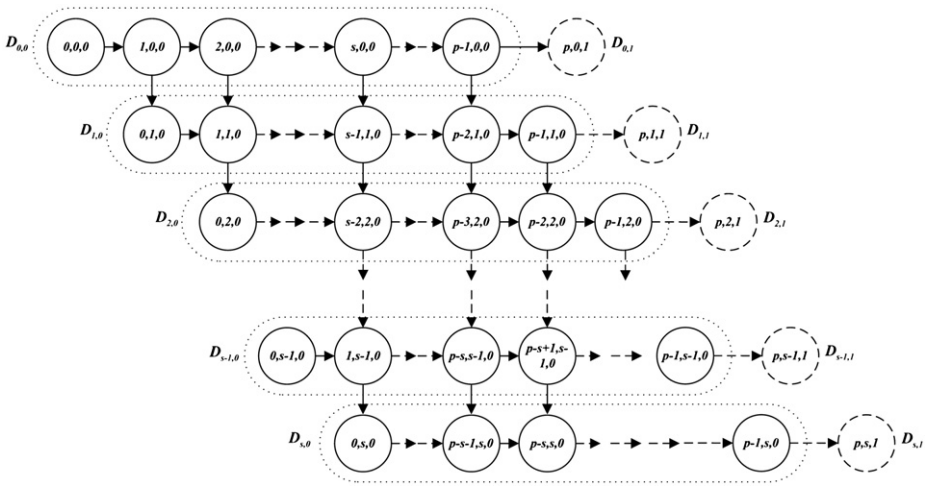

Fig. 1. Partial transition diagram of the system for classes in the form of $D_{j,0}$.

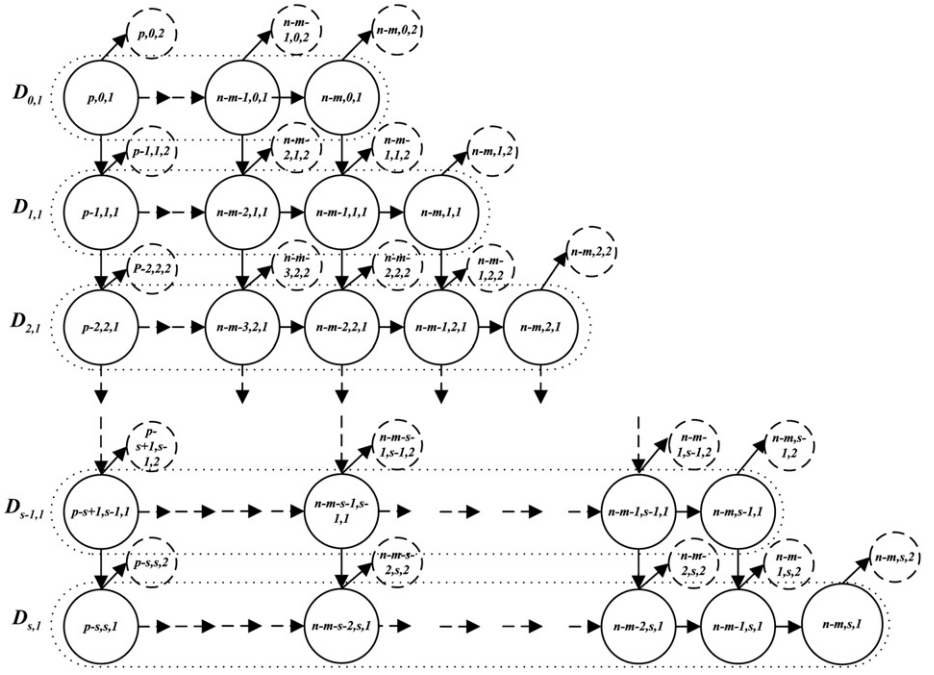

Fig. 2. Partial transition diagram of the system for classes in the form of $D_{j,1}$.

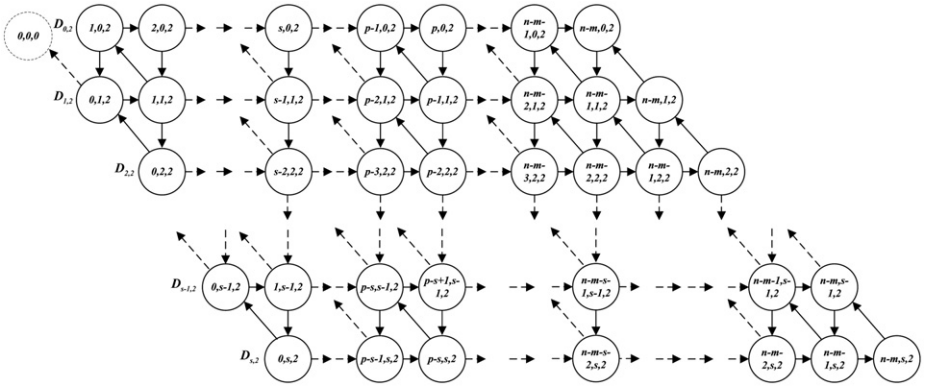

Fig. 3. Partial transition diagram of the system for classes in the form of $D_{j,2}$.

conditions as

$$N=(s+1)\left(2(n-m+1)+\frac{s}{2}\right)-1,\quad p\geq s \text{ and } p\neq 1. \tag{4}$$

$$N=2(s+1)(n-m+1)+sp-\frac{p^2}{2}+\frac{p}{2}-1,\quad p<s \text{ and } p\neq 1. \tag{5}$$

$$N=2(s+1)(n-m+1)-1,\quad p=1. \tag{6}$$

Now that the system's state space and transition diagram are defined, we need to find all possible transitions between these states in the system.

2.3. Transition rates

Due to the large number of reachable states in the system, a systematic approach is needed to find the transition rates in the system. As shown in Figs. 1–3, there are four types of transitions in the system which are forward transitions ($\rightarrow$), backward transitions ($\nwarrow$), vertical transitions ($\downarrow$), and diagonal transitions ($\nearrow$). A forward transition occurs when one active component fails, a backward transition occurs when the repair of one failed component is completed, a vertical

transition occurs when a cold standby component replaces a failed active component, and a diagonal transition occurs when maintenance initiation setup is completed. Now, the respective non-zero failure rates $\lambda_{(i,j,z)}$, repair rates $\mu_{(i,j,z)}$, switching rates $\delta_{(i,j,z)}$, and maintenance initiation setup rates $\varepsilon_{(i,j,z)}$ of each state (i,j,z) in the system can be defined as follows:

$$\lambda_{(i,j,z)} = \begin{cases} (n-i)\lambda, & (i,j,z) \in D, \quad i \neq n-m \\ (n-i-m)\lambda, & (i,j,z) \in \overline{D}, \quad i \neq n-m \end{cases}, \tag{7}$$

$$\mu_{(i,j,z)} = \min(j,r)\mu, \quad (i,j,z) \in D, \quad z = 2, \tag{8}$$

$$\delta_{(i,j,z)} = \min(i,s-j)\delta, \quad (i,j,z) \in D, \quad j < s, \tag{9}$$

$$\varepsilon_{(i,j,z)} = \varepsilon, \quad (i,j,z) \in D, \quad z = 1. \tag{10}$$

It can be shown from Eqs. (7)–(10) that each state can have at most one forward, one backward, one vertical, and one diagonal transitions. In order to be able to generate the transition matrix of the system, connected states after each transition should be systematically defined. These connected states after each forward ($F_{(i,j,z)}$), backward ($B_{(i,j,z)}$), vertical ($V_{(i,j,z)}$), and diagonal ($I_{(i,j,z)}$) transitions for state (i,j,z) can be defined as follows:

$$F_{(i,j,z)} = \begin{cases} (i+1,j,z), & (i,j,z) \in D, \quad i \neq n-m, \quad (i,j,z) \neq (p-1,j,1) \\ (p,j,1), & (i,j,z) = (p-1,j,0) \end{cases}, \tag{11}$$

$$B_{(i,j,z)} = \begin{cases} (i,j-1,z), & (i,j,z) \in D, \quad j \geq 1, \quad z = 2, \quad i+j \neq 1 \\ (0,0,0), & (i,j,z) = (0,1,2) \end{cases}, \tag{12}$$

$$V_{(i,j,z)} = (i-1,j+1,z), \quad (i,j,z) \in D, \quad i \geq 1, \quad j < s, \tag{13}$$

$$I_{(i,j,z)} = (i,j,2), \quad (i,j,z) \in D, \quad z = 1. \tag{14}$$

Although the transition rates between states can be defined based on the above equations, finding the complete transition matrix of the system is complicated and burdensome. Here, using Eqs. (11)–(14), we introduce a systematic approach which can be used to automatically construct the transition matrix of the problem. Rather than constructing the complete transition matrix for the system by searching all transitions, smaller matrices showing the transitions between two classes of states are generated first. Here, we first define partial binary matrices for different types of transitions among two classes of states and finally combine these matrices to get the completed infinitesimal generator of the problem. These types of binary transition matrices are denoted by $Q^1_{A,A'}$, $Q^2_{A,A'}$, $Q^3_{A,A'}$, and $Q^4_{A,A'}$, where A and A' are two classes of states from D. The entry of one in the ith row and the jth column of matrix $Q^f_{A,A'}$, $f \in (1,2,3,4)$, indicates that there is a transition of type f from the ith state in A to the jth state in A'. We will show that these matrices have a unique and simple structure which can simplify constructing the system's infinitesimal generator. The forward binary transition matrix ($Q^1_{A,A'}$) can be expressed as

$$Q^1_{A,A} = \begin{bmatrix} 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}, \quad \{A = (D_{j,z}) \mid 0 \leq j \leq s, 0 \leq z \leq 2\}, \tag{15}$$

$$Q^1_{A,A'} = \begin{bmatrix} & & \overbrace{\quad}^{\min(j+1,p+1)} & & \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & \dots & 0 \end{bmatrix}, \quad \{A = (D_{j,0}), \quad A' = (D_{j,1}) \mid 0 \leq j \leq s\}. \tag{16}$$

Eqs. (15)–(16) verify that the forward transitions occur within the states in the same class (Eq. (15)) and between the last state in all classes of the form $D_{j,0}$ and their corresponding states in classes of the form $D_{j,1}$ (Eq. (16)). The binary backward transition matrix ($Q^2_{A,A'}$) for two classes of states A and A' is defined as follows:

$$Q^2_{A,A'} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}, \quad \{A = (D_{j,2}), \quad A' = (D_{j-1,2}) \mid 2 \leq j \leq s\}, \tag{17}$$

$$Q^2_{A,A'}=\begin{bmatrix}0&0&\dots&0&0\\1&0&\dots&0&0\\0&1&\dots&0&0\\\vdots&\vdots&\dots&\vdots&\vdots\\0&0&\dots&0&1\end{bmatrix},\quad \{A=(D_{1,2}),\quad A'=(D_{0,2})\}. \tag{18}$$

$$Q^2_{A,A'}=\begin{bmatrix}1&0&\dots&0&0\\0&0&\dots&0&0\\\vdots&\vdots&\vdots&\vdots&\vdots\\0&0&\dots&0&0\end{bmatrix},\quad \{A=(D_{1,2}),\quad A'=(D_{0,0})\}. \tag{19}$$

As shown in Eqs. (17)–(19), backward transitions occur for states in classes of the form $D_{j,2}$. The binary vertical transition matrix ($Q^3_{A,A'}$) for two classes of states A and A' is defined as follows:

$$Q^3_{A,A'}=\begin{bmatrix}0&0&\dots&0&0\\1&0&\dots&0&0\\0&1&\dots&0&0\\\vdots&\vdots&\dots&\vdots&\vdots\\0&0&\dots&1&0\end{bmatrix},\quad \left\{\begin{matrix}A=(D_{j,z}),\quad A'=(D_{j+1,z})|0\le j\le s-1,\quad z\neq 1,\quad A\neq(D_{0,2})\\A=(D_{j,1}),\quad A'=(D_{j+1,1})|0\le j\le s-1,\quad j\ge p\end{matrix}\right\}, \tag{20}$$

$$Q^3_{A,A'}=\begin{bmatrix}1&0&\dots&0&0\\0&1&\dots&0&0\\0&0&\dots&0&0\\\vdots&\vdots&\dots&\vdots&\vdots\\0&0&\dots&1&0\end{bmatrix},\quad \left\{\begin{matrix}A=(D_{j,1}),\quad A'=(D_{j+1,1})|0\le j\le s-1,\quad j<p\\A=(D_{0,2}),\quad A'=(D_{1,2})\end{matrix}\right\}, \tag{21}$$

As shown in Eqs. (20)–(21), vertical transitions occur for all states with at least one non-failed cold standby component and one failed active component. The binary diagonal transition matrix ($Q^4_{A,A'}$) for two classes of states A and A' is defined as follows:

$$Q^4_{A,A'}=\begin{bmatrix}&\overbrace{}^{p-j+1}&&&&\\\dots&1&0&\dots&0&\dots\\\dots&0&1&\dots&0&\dots\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\\dots&0&0&\dots&1&0\\&&&&&\overbrace{}^{n-m+1}\\\dots&0&0&\dots&0&1\end{bmatrix},\quad \{A=(D_{j,1}),\quad A'=(D_{j,2})|1\le j\le \min(s,p)\}, \tag{22}$$

$$Q^4_{A,A'}=\begin{bmatrix}1&0&\dots&0&0\\\vdots&\vdots&\vdots&\vdots&\vdots\\0&0&\dots&0&0\\0&0&\dots&1&0\\0&0&\dots&0&1\end{bmatrix},\quad \{A=(D_{j,1}),\quad A'=(D_{j,2})|p\le j\le s\}, \tag{23}$$

$$Q^4_{A,A'}=\begin{bmatrix}&\overbrace{}^{p}&&&\\\dots&1&0&0&0\\&&\overbrace{}^{p+1}&&\\\dots&0&1&0&0\\\vdots&\vdots&\vdots&\vdots&\vdots\\\dots&0&0&\dots&0\\&&&&\overbrace{}^{n-m}\\\dots&0&0&0&1\end{bmatrix},\quad \{A=(D_{0,1}),\quad A'=(D_{0,2})\}. \tag{24}$$

As shown in Eqs. (22)–(24), diagonal transitions occur for all states in classes of the form $D_{j,1}$, $1\le j\le s$. To construct the complete transition matrix of the system, we will define binary forward, backward, vertical, and diagonal transition

matrices of the system as follows:

$$Q_D^f=\begin{bmatrix} Q^f_{D_{0,0},D_{0,0}} & Q^f_{D_{0,0},D_{1,0}} & \cdots & Q^f_{D_{0,0},D_{j,z}} & \cdots & Q^f_{D_{0,0},D_{s-1,2}} & Q^f_{D_{0,0},D_{s,2}} \\ Q^f_{D_{1,0},D_{0,0}} & Q^1_{D_{1,0},D_{1,0}} & \cdots & Q^f_{D_{1,0},D_{j,z}} & \cdots & Q^f_{D_{1,0},D_{s-1,2}} & Q^f_{D_{1,0},D_{s,2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Q^f_{D_{s-1,2},D_{0,0}} & Q^f_{D_{s-1,2},D_{1,0}} & \cdots & Q^f_{D_{s-1,2},D_{j,z}} & \cdots & Q^f_{D_{s-1,2},D_{s-1,2}} & Q^f_{D_{s-1,1},D_{s,2}} \\ Q^f_{D_{s,2},D_{0,0}} & Q^f_{D_{s,2},D_{1,0}} & \cdots & Q^f_{D_{s,2},D_{j,z}} & \cdots & Q^f_{D_{s,2},D_{s-1,2}} & Q^f_{D_{s,2},D_{s,2}} \end{bmatrix}, \quad f\in(1,2,3,4). \tag{25}$$

These binary matrices are generated by combining all the same type binary transition matrices defined in Eqs. (15)–(24). Based on the above definitions, Q_D^1, Q_D^2, Q_D^3, and Q_D^4 are square binary matrices ($N\times N$) that show the relationships among all states in the system. If we replace all unit values in these matrices with their corresponding transition rates (obtained from Eqs. (7)–(10)), the infinitesimal generator of the problem can be constructed. To do this, we define square matrices $\overline{F}$, $\overline{B}$, $\overline{V}$, and $\overline{I}$, in the sense that the elements in the jth row of all columns refer to the corresponding failure rate, repair rate, switching rate, and maintenance setup rate of the jth state in D, respectively. The initial infinitesimal generator of the problem can now be created as

$$Q_D^0=Q_D^1\bullet\overline{F}+Q_D^2\bullet\overline{B}+Q_D^3\bullet\overline{V}+Q_D^4\bullet\overline{I}, \tag{26}$$

where • denotes the scalar element-by-element multiplication operator. Since the above matrix (Eq. (26)) is the initial form of the infinitesimal generator of the model, we manually set all diagonal elements of matrix Q, so that the sum of all elements in each row becomes zero. The diagonal elements of this matrix can be calculated as $-diag(Q_D^0\times[1])$, where [1] is a column vector with all elements equal to 1. Now the system of linear equations, $\pi\ Q=0$, can be employed to find the steady state measures of the system considering the normalization condition, $\pi\times[1]=1$.

It is important to note that since all the expressions used in the optimization model described in Section 4 including the objective function and constraints are directly or indirectly derived from the steady state solutions of the system, the time required to calculate the objective function and constraints associated with a single solution point of the optimization model is related to the size of the system of linear equations, $\pi Q=0$. In Eqs. (4)–(6), we calculated the number of states (N) under three circumstances. Depending on the type of the method used to solve this system of equations (such as Cramer's rule, Gaussian Elimination, etc.), the time required to find the solutions of this system of linear equations may vary.

3. Performance measures of the system

In the following, we introduce some important steady state performance measures, which will be used in the design and maintenance optimization model. Some of these measures have been defined in Moghaddass and Zuo (2011).

a. Steady state probability that the system is in the health state w:

$$\pi'_w=\Pr_{t\to\infty}(d_{w+1}>G(t)\geq d_w)=\sum\nolimits_{(i,j,z)\in A}\pi_{(i,j,z)},\quad A=\{(i,j,z)\,|\,n-O_{w+1}<i\leq n-O_w\}. \tag{27}$$

b. Steady state availability (the probability of satisfying the minimum required demand):

$$SSA=\Pr_{t\to\infty}(G(t)\geq d_{M-1})=\sum\nolimits_{(i,j,z)\in A}\pi_{(i,j,z)},\quad A=\{(i,j,z)\,|\,i\leq n-O_{M-1}\}. \tag{28}$$

c. Mean time between system failures (the average time between two consecutive failures):

$$MTBF=\frac{1}{Failure\ Frequency}=\frac{1}{\sum_{(i,j,z)\in A}\pi_{(i,j,z)}\lambda_{(i,j,z)}},\quad A=\{(i,j,z)\,|\,i=n-O_{M-1}\} \tag{29}$$

d. Mean downtime of the system in each cycle (MDT):

$$MDT=(1-SSA)MTBF. \tag{30}$$

e. Mean repair time in each cycle (MRT):

$$MRT=\left[\sum\nolimits_{(i,j,z)\in A}\pi_{(i,j,z)}\min(j,r)\right]MTBF,\quad A=\{(i,j,z)\,|\,z=2\}. \tag{31}$$

f. Mean hiring time in each cycle (MHT) (the average repairman-hours in each cycle):

$$MHT=\left[\sum\nolimits_{(i,j,z)\in A}\pi_{(i,j,z)}r\right]MTBF,\quad A=\{(i,j,z)\,|\,z=2\}. \tag{32}$$

g. Mean number of maintenance setups in each cycle (MNM):

$$\text{Maintenace Frequency } (MF) = \left(\sum_{j=0}^{s}(\pi_{(p-1,j,0)}\lambda_{(p-1,j,0))}\right) \rightarrow MNM = MTBF \times MF \tag{33}$$

h. Mean non-functional time of cold standby components in each cycle (MRN_1):

$$MRN_1 = \left[\sum_{(i,j,z)\in D}\pi_{(i,j,z)}(s)\right]MTBF \tag{34}$$

i. Mean non-functional time of active components in each cycle (MRN_2):

$$MRN_2 = \left[\sum_{(i,j,z)\in D}\pi_{(i,j,z)}(i)\right]MTBF. \tag{35}$$

j. Mean additional operation time of active redundant components in each cycle (MOT):

$$MOT = \left[\sum_{(i,j,z)\in D}\pi_{(i,j,z)}\max(0,(n-i-O_{M-1}))\right]MTBF. \tag{36}$$

k. Mean sojourn time in the health state w in each cycle:

$$MST_w = \pi'_w \times MTBF. \tag{37}$$

Based on the above measures, we define the cost of additional operation in each cycle as $c_1 \times MOT$, the cost of non-functionality for cold standby components in each cycle as $c_2^{(1)} \times MRN_1$, the cost of non-functionality for active components in each cycle as $c_2^{(2)} \times MRN_2$, the cost of server activations in each cycle as $c_3 \times MNM$, the cost of hiring repair facilities (repairman) in each cycle as $c_4 \times MHT$, the cost of repair in each cycle as $c_5 \times MRT$, the loss per system downtime in each cycle as $c'_M \times MDT$, and the profit per being in health condition w in each cycle as $c'_w \times MST_w$. As a control measure, we can also calculate the mean duration for maintenance setup (MMS) in each cycle as

$$MMS = \left[\sum_{(i,j,z)\in A}\pi_{(i,j,z)}\right]MTBF, \quad A = \{(i,j,z)|z=1\}, \tag{38}$$

The above measure demonstrates the portion of each cycle where the system is waiting to be repaired. Also, the average performance level of the multi-state system can be expressed as

$$G = \sum_{(i,j,z)\in D}\pi_{(i,j,z)}(n-i)E. \tag{39}$$

4. Optimization model

The unit profit function can be maximized using the following optimization model:

Maximize Unit Profit

$$= \frac{\sum_{w=1}^{M-1}c'_w \times MST_w - c_1 \times MOT - c_2^{(1)} \times MRN_1 - c_2^{(2)} \times MRN_2 - c_3 \times MNM - c_4 \times MHT - c_5 \times MRT - c'_M \times MDT}{MTBF}, \tag{40}$$

subject to

$$SSA \geq b_{SSA}, \quad c_6(n+s) \leq b_1, \quad O_{M-1} \leq n \leq b_2, \quad 1 \leq s \leq b_3, \quad 1 \leq r \leq s, \quad 1 \leq p \leq n-O_{M-1}+1, \quad n,s,r,p : \text{integer},$$

where b_{SSA} is the minimum required availability for the system. The lower bound of n is O_{M-1}, since at least O_{M-1} working components are needed for the healthy operation of the system (to satisfy the minimum required demand). The upper bound for n can be determined based on factors such as budget and space. The upper and lower bounds of s can also be determined based on budget, space, and other considerations; however there should be at least one standby component. The lower bound of r is 1, unless specified otherwise. The maximum number of failed cold standby components is s, therefore the upper bound for r is s, which is the maximum number of possible failed components which can be simultaneously repaired in the system. The lower bound for the maintenance activation point is equal to one, unless specified otherwise. The latest we can send the repair group to perform maintenance is when the system has failed ($p=n-O_{M-1}+1$). It is important to note that due to the integer types of all decision variables and the bound constraints on all decision variables, the optimization model described in Eq. (40) has a finite number of solution points. The number of solution points for this model is as

$$\sum_{n=O_{M-1}}^{b_2}\sum_{s=1,(n+s)\leq\frac{b_1}{c_6}}^{b_3}(s(n-O_{M-1}+1)).$$

Based on this limited number of solution points, a complete enumeration could be used to find the global optimum solution of the problem.

5. Numerical example

The optimization problem described in this paper is a constrained non-linear integer programming model with a limited number of solution points. However, depending on the bounds given for decision variables, complete enumeration (exhaustive search) may take a huge amount of time. That is why any kind of metaheauristic, such as GA, can be used to find the optimal solution in a shorter time period. Although evaluating the performance and efficiency of possible evolutionary algorithms to be used to find the optimal solution of our model is out of the scope of this paper, it is to be noted that this type of optimization problem is generally an offline procedure, that is, decision making on design and maintenance variables is usually not online or repetitive. This means that the time required to find the optimum solution is not critical and also is not of high concern for such type of problems. Therefore, it is the decision makers' choice to decide which method to use for finding the optimal solution, depending on the size of the solution space and the required level of accuracy.

In this paper, Genetic Algorithm (GA) which is a widely used meta-heuristic approach for solving large optimization problems is employed due to its flexibility in representing design variables in a discrete form and its good global optimization capability, especially for similar problems in reliability and maintenance optimization (Levitin and Lisnianski, 2000; Liu and Huang, 2010; Machani and Nourelafath, 2010). Also, considering the limited number of solution points for the problem, GA can provide reasonable results, if a good initial population is selected.

5.1. Numerical example: power generation system

A power generation system needs to supply different customers with different demands (Levitin, 2005). Consider a power generation system with n s-independent power generators with binary-states. The cumulative power generated from all available active operating power generators is the total power output of this type of power generator system. Therefore, $G(t)=\Sigma_{i=1}^{n}G_i(t)$. The nominal capacity of each power generator is 40 MW ($E=40$). There are 4 levels of demand which are $d_1=640$ MW, $d_2=480$ MW, $d_3=320$ MW, and $d_4=160$ MW. Therefore, the minimum required power is 160 MW, which means that if the cumulative power output of all of the available generators is not sufficient to meet this demand, then the system is considered unavailable. Also, the system may generate more profit if a higher level of demand is met. The profit that the system will generate per unit time if it satisfies the demand levels 1 to 4 are $c'_1=\$1400$, $c'_2=\$1200$, $c'_3=\$1000$, and $c'_4=\$800$. The cost of system downtime (producing less than 160 MW) per unit time is $c'_M=\$10{,}000$.

There is an activation cost and maintenance initiation setup time for repair; also using each repair facility entails operational expenses. There is, on the other hand, a cost for a power generator that remains non-functional. Hence having more cold standby components may incur additional nonoperational expenses, but system downtime can cost even more, whenever the minimum level of demand cannot be satisfied. Moreover, there is a switching time for each cold standby generator to become an active generator. From maintenance point of view, we need to find out how many repair facilities should be sent for repair and when is the best time to send them. Other inputs of the numerical example are $\lambda=1/(20160\text{ min})$, $\mu=1/(2880\text{ min})$, $\delta=1/(1440\text{ min})$, and $\varepsilon=1/(10080\text{ min})$, $\varepsilon=95\%$, $c_1=\$80$, $c_2^{(1)}=\$10$, $c_2^{(2)}=\$15$, $c_3=\$20{,}000$, $c_4=\$10$, $c_5=\$10$, $c_6=\$90{,}000$, $b_1=\$2{,}000{,}000$, $b_2=20$ units, $b_3=20$ units, and $m=0$.

The corresponding stochastic process is formulated and the optimization model is programmed using Matlab R2008. We used a computer with 2.00 GHZ AMD (tm) processor and 3.00 GB RAM under Windows XP operating system. The population data type was set at double and the population size at 20. The stopping criteria were set at 100 generations, 100 stall generations, and 10^{-9} as the function tolerance. Other setups were set as Matlab default. The optimum values found with GA are $n^*=8$, $s^*=14$, $r^*=2$, and $p^*=1$. Based on this optimal solution, from design point of view, there should be 8 active generators and 14 cold standby generators in the system. As for the optimal maintenance strategy, whenever an active generator fails, 2 repair facilities should be sent for repair. Other important performance measures of this system with these optimal solutions are shown in Table 1.

Table 1
Performance measures of the optimal design.

Measures	Opt. value	Measures	Opt. value	Measures	Opt. value
Unit profit	$ 413.56	MDT	4 days	MST1	0
SSA_M	99.51%	MRT	863.1 days	MST2	0
CPU time	4.47 min	MHT	1031.2 days	MST3	459.34 days
MTBF	815.5 days	MNM	34.27	MST4	352.16 days

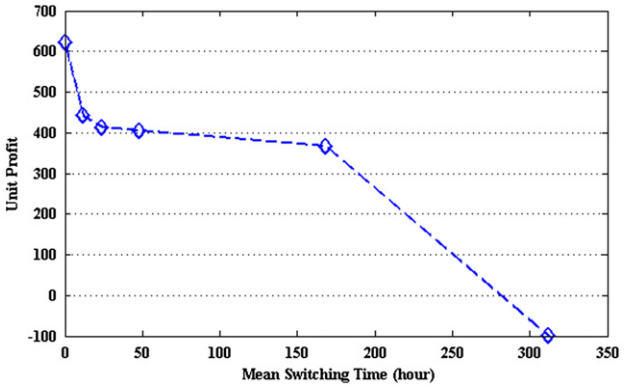

Fig. 4. Effect of the mean switching time on the optimal unit profit of the system.

5.2. Discussions of the optimum solution

After the optimal solution is obtained, the maintenance team can plan for the preventive condition-based maintenance activities. It is clear that changing parameters such as the maintenance initiation setup rate may change the optimum solution, i.e. shorter maintenance initiation setup time may result in sending repair facilities later (when more than one active generator has failed, i.e. $p^* > 1$). The cycle time of the system or MTBF is 815.5 days which means that on average, every 815.5 day, the power generation system will be down for 4 days and will be unable to satisfy the minimum demand of 160 MW. This short downtime period complies with the obtained availability of 99.51%.

Using the information provided in Table 1 enables the maintenance team to manage repair facilities accordingly. It should be noted that at any time, if any of the associated costs of the system changes, a new optimization model is to be constructed to find the new optimum strategy. Due to the simple structure of the model, decision makers can easily perform sensitivity analysis and choose the best possible solutions under different circumstances. Also different types of constraints such as the minimum length for maintenance period can be added to the model. Rather than long-run analysis, the short term behavior of the system can also be analyzed by employing transient solutions.

To show the advantage of this model over similar work in the literature where maintenance activation time and switching time for standby components were assumed negligible, as an example, we show the importance of the former factor by presenting its effect on the optimal unit cost (profit) of the system. Fig. 4 illustrates the trend of the optimal unit profit of the system versus the mean switching time.

The result shown in Fig. 4 verifies that the system will earn lower benefit if the switching time for cold standby components increases, i.e. if it exceeds a certain limit, then the system become uneconomical.

6. Conclusion and future work

In this work, we have formulated the trade-off between the optimal design of a repairable multi-state system with binary-state components and maintenance strategies. Decision variables such as the number of active components, the number of cold standby components, the number of repair facilities, and the maintenance activation point are considered in the optimization model. A systematic approach to find the system's state space, transition matrix, and performance measures of such system is introduced. The corresponding cost-effective optimization model is developed to maximize the unit profit of the system so that the minimal availability requirement for the system is met. The application of this method to multi-state systems with multi-state components and non-exponential failure and repair rates is to be explored in future work.

Acknowledgment

This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Comments and suggestions from reviewers and the editor are very much appreciated.

References

Amari, S.V., Dill, G., 2009. A new method for reliability analysis of standby systems. In: Proceedings of Annual Reliability and Maintainability Symposium, art. no. 4914713, pp. 417–422.

Amari, S.V., Pham, H., 2007. A novel approach for optimal cost-effective design of complex repairable systems. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 37 (3), 406–415.

Chiang, J.H., Yuan, J., 2001. Optimal maintenance policy for a Markovian system under periodic inspection. Reliability Engineering & System Safety 71 (2), 165–172.
Fawzi, B.B., Hawkes, A.G., 1991. Availability of an r-out-of-n system with spares and repairs. Journal of Applied Probability 28 (2), 397–408.
Gurov, S.V., Utkin, L.V., Shubinsky, I.B., 1995. Optimal reliability allocation of redundant units and repair facilities by arbitrary failure and repair distribution. Microelectronics Reliability 35 (12), 1451–1460.
Huang, J., Zuo, M.J., Wu, Y., 2000. Generalized multi-state k-out-of-n:G systems. IEEE Transactions on Reliability 49 (1), 105–111.
Krishnamoorthy, A., Ushakumari, P.V., 2001. k-out-of-n: G system with repair: the D-policy. Computers and Operations Research 28, 973–981.
Krishnamoorthy, A., Ushakumari, P.V., Lakshmi, B., 2002. k-out-of-n system with repair: the N policy. Asia Pacific Journal of Operational Research 19, 47–61.
Lad, B.K., Kulkarni, M.S., Misra, K.B., 2008. Optimal reliability design of a system. Handbook of Performability Engineering. Springer, London.
Levitin, G., 2005. Introduction to multi-state systems. Universal Generating Function in Reliability Analysis and Optimization, Springer Series in Reliability Engineering. Springer, London, pp. 67–98 (chapter 3).
Levitin, G., Lisnianski, A., 2000. Optimization of imperfect preventive maintenance for multi-state systems. Reliability Engineering & System Safety 67 (2), 193–203.
Lisnianski, A., Levitin, G., 2003. Multi-state System Reliability Assessment, Optimization and Application. World Scientific Publishing Co., New York.
Liu, Yu, Huang, Hong-Zhong, 2010. Optimal selective maintenance strategy for multi-state systems under imperfect maintenance. IEEE Transactions on Reliability 59 (2), 356–367.
Machani, M., Nourelafath, M., 2010. A genetic algorithm for integrated production and preventive maintenance planning in multi-state systems. In: Proceedings of the Eighth International Conference of Modeling and Simulation—Mosim'10, Evaluation and optimization of innovative production systems of goods and services, Hammamet, Tunisia.
Misra, K.B., 1974. Reliability design of a maintained system. Microelectronics Reliability 13 (6), 495–500.
Moghaddass, R., Zuo, M.J., 2011. Optimal design of repairable k-out-of-n systems considering maintenance. In: Proceedings of the Annual Reliability and Maintainability Symposium, Orlando, USA.
Moghaddass, R., Zuo, M.J., Qu, J., 2011a. Reliability and Availability Analysis of a Repairable K-out-of-N:G System with R Repairmen Subject To Shut-Off Rules. IEEE Transactions on Reliability 60 (3), 666–858.
Moghaddass, R., Zuo, M.J., Wang, W., 2011b. Availability of a general k-out-of-n:G system with non-identical components considering shut-off rules using quasi-birth–death process. Reliability Engineering & System Safety 96 (4), 489–496.
Mohamed, A.A., 1992. Optimization techniques for system reliability: a review. Reliability Engineering & System Safety 35 (2), 137–146.
Pham, H., 1992. On the optimal design of k-out-of-n subsystems. IEEE Transactions on Reliability 41 (4), 572–574.
Sasaki, M., Kaburaki, S., Yanagi, S., 1977. System availability and optimum spare units. IEEE Transactions on Reliability R-26 (3), 182–187.
Sharma, U., Misra, K.B., 1988. Optimal availability design of a maintained system. Reliability Engineering and System Safety 20 (2), 147–159.
Sleptchenko, A., Van der Heijden, M.C., Van Harten, A., 2003. Trade-off between inventory and repair capacity in spare part networks. The Journal of the Operational Research Society 54 (3), 263–272.
De Smidt-Destombes, K.S., Van der Heijden, M.C., Van Harten, A., 2009. Joint optimization of spare part inventory, maintenance frequency and repair capacity for k-out-of-N systems. International Journal of Production Economics 118 (1), 260–268.
De Smidt-Destombes, K.S., Van der Heijden, M.C., 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.S., Van der Heijden, M.C., 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 & System Safety 83 (3), 287–300.
Srivastava, V.K., Fahim, A., 1988. K-out-of-m system availability with minimum cost allocation of spares. IEEE Transactions on Reliability 37 (3), 287–292.
Ushakumari, P.V., Krishnamoorthy, A., 2004. k-out-of-n system with repair: the max (N, T) policy. Performance Evaluation 57, 221–234.
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.
Wang, K., 1993. Cost analysis of the M/M/R machine-repair problem with mixed standby spares. Microelectronics and Reliability 33 (9), 1293–1301.