Available online at www.sciencedirect.com

SCIENCE DIRECT®

Reliability Engineering and System Safety 91 (2006) 20–35

RELIABILITY ENGINEERING & SYSTEM SAFETY

www.elsevier.com/locate/ress

Risk-informed optimisation of railway tracks inspection and maintenance procedures

Luca Podofillini[a], Enrico Zio[a,*], Jørn Vatn[b]

[a]Department of Nuclear Engineering, Polytechnic of Milan, via Ponzio 34/3, 20133 Milan, Italy
[b]Department of Production and Quality Engineering, NTNU, 7491 Trondheim, Norway

Received 30 June 2004; accepted 13 November 2004
Available online 21 January 2005

Abstract

Nowadays, efforts are being made by the railway industry for the application of reliability-based and risk-informed approaches to maintenance optimisation of railway infrastructures, with the aim of reducing the operation and maintenance expenditures while still assuring high safety standards.

In particular, in this paper, we address the use of ultrasonic inspection cars and develop a methodology for the determination of an optimal strategy for their use. A model is developed to calculate the risks and costs associated with an inspection strategy, giving credit to the realistic issues of the rail failure process and including the actual inspection and maintenance procedures followed by the railway company.

A multi-objective optimisation viewpoint is adopted in an effort to optimise inspection and maintenance procedures with respect to both economical and safety-related aspects. More precisely, the objective functions here considered are such to drive the search towards solutions characterized by low expenditures and low derailment probability. The optimisation is performed by means of a genetic algorithm. The work has been carried out within a study of the Norwegian National Rail Administration (Jernbaneverket).

1. Introduction

In the past, railway maintenance procedures have been traditionally planned based on the knowledge and experience of each company, accumulated over many decades of operation, but without any kind of reliability- or risk-based approaches and with the major goal of providing a high level of safety to the infrastructures, without much concern over the economical issues [1].

However, nowadays, the railway business has to compete with alternative forms of transportation so that reducing operational expenditures has become a major task for competitiveness. To this aim, efforts are being made for the application of reliability-based and risk-informed approaches to maintenance optimisation of railway infrastructures. The underlying idea is to reduce the operation and maintenance expenditures while still assuring high safety standards [1–4].

A major threat to the safe operation of a railway system is represented by the presence of defects such as cracks and of geometry failures like track misalignments. Defects are initiated within the rail due to fatigue and other failure mechanisms. As the rail operation proceeds, defects can worsen if no recovery action is undertaken and, in the end, they may develop to complete rail breakage, which is a major cause for train derailment.

Monitoring cars are used to detect the defects before they develop to rail breakage. In modern railway operation there are two types of monitoring cars:

- Ordinary measurement cars measuring rail geometry and surface deterioration
- Ultrasonic inspection (USI) cars measuring rail breakage and internal cracks.

In this paper, we will focus on the use of ultrasonic inspection cars and develop a methodology for the determination of an optimal strategy for their use.

* Corresponding author. Tel.: +39 02 2399 6340; fax: +39 02 2399 6309.
E-mail address: enrico.zio@polimi.it (E. Zio).

0951-8320/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2004.11.009

The work has been carried out within a study of the Norwegian National Rail Administration (Jernbaneverket, JBV). The problem addressed is twofold. On one side, the verification of the goodness of an inspection strategy must rely on a comprehensive model to calculate the associated risks and costs. Such model should account, as closely as possible, for the realistic issues of the rail failure process and include the actual inspection and maintenance procedures followed by the railway company. On the other side, different maintenance strategies must be searched and evaluated within an optimisation scheme addressing both the above mentioned economical and safety aspects.

1.1. The quantitative system model

The model here developed for estimating the rail failure probability is based on the concept of 'P–F interval', i.e. the interval between the time a potential failure can be observed and that at which the failure actually occurs [5]. The basic idea under this concept is that failure is regarded as a two-stage process. First, at some time a defect in the system becomes detectable, then, after some delay-time, the system fails due to the degeneration of the defect. An exhaustive review of the models based on the P–F interval concept can be found in Ref. [5]. From our point of view, the P–F interval represents the time available to detect the presence of a defect before it causes the actual failure. A common hypothesis is that the probability of defect detection does not depend on the time since the defect originated. This does not seem a realistic hypothesis since the probability of detecting a defect should depend on whether the defect is newly born or it is well developed [6].

A further complication of the model is the need of including the actual test/maintenance procedures followed in the practice of railway operation. In this respect, different recovery actions may follow the detection of a defect, based on a categorization of its severity. For example, a defect may not necessarily be fixed when detected, but, if its criticality allows it, a decision may be taken to continue in normal operation and wait for a more suitable time to fix it. For instance, a common 'opportunistic' procedure is to wait until a given number of defects have been detected and then fix them all at once. Maintenance procedures must then be introduced to define the criticality criteria of the defect and, correspondingly, when to repair them.

In this paper, we propose a non-homogeneous Markov model for determining the failure probability of a rail section under periodic inspection, giving explicit credit to the degradation process and the different inspection and maintenance procedures. The present work extends a preliminary model, developed by one of the authors within the JBV study [7], by incorporating deterministic state transitions in correspondence of the times of the inspection and maintenance. The general features of the modelling approach are presented in Ref. [9], where close formulae, based on inverse and exponential matrices, are derived for some reliability measures on small-size problems. When maintenance procedures are included in the model, the complexity of the problem increases significantly and such formulae become cumbersome. To overcome this, in this paper we resort to a simple numerical procedure for the integration of the equations governing the process dynamics.

1.2. The optimization approach

Concerning the optimisation of the inspection strategy, classical gradient-descent based methods, dynamic programming, integer programming, mixed-integer and non-linear programming could be employed [10]. However, these optimisation methods typically resort to simplified models to provide explicit expressions for the objective function to be optimised and its derivatives. Unfortunately, the complex behaviour of modern industrial plants and systems can hardly be captured by explicit analytical models so that the objective function, and its dependence on the control and decision variables, are embedded into intricate computer codes. This poses severe limitations to the applicability of the above mentioned classical optimisation methods. In these situations, modern numerical search algorithms, such as Genetic Algorithms (GAs), have shown great potentialities [11–25]. Genetic Algorithms are numerical search tools, which operate according to procedures that resemble the principles of natural selection and genetics [11,12]. Because of their flexibility, global perspective and non reliance on differential information for their operation, GAs have been successfully used in a wide variety of problems in several areas of engineering and life science [13–19].

In an effort to optimise inspection and maintenance procedures with respect to both economical and safety-related aspects, we adopt a multi-objective optimisation viewpoint [20–25]. More precisely, the objective functions here considered are such to drive the search towards solutions characterized by low expenditures and low derailment probability.

The paper is structured as follows. In Section 2, we synthesize the basic model used within the JBV study [7] describing the implications that inspections/maintenance strategies have on the safety and on the economics of the railway infrastructure. This model defines the general decisional framework and requires to compute the probability that a crack is not detected by USI. In Section 3, we describe the general approach for incorporating inspections into a Markov modelling framework [9]. In Section 4 we present the specific model for evaluating the probability of interest. In Section 5, we briefly present the fundamentals governing the standard genetic algorithm search procedure, and its multiobjective extension. Section 6 reports the results of the optimisation. Some concluding remarks are given in the last Section.

2. Problem description

A conceptual risk model of the railway infrastructure is shown in Fig. 1. The Figure sketches the different barriers that prevent an initiated failure from leading to derailment. The barriers are:

- *Rail quality*: The rails are constructed to have a strength, under nominal operation loads, that in principle should ensure an infinite life length. Unfortunately, high axle loads, wheel failures and superstructure failures can compromise this barrier significantly.
- *Ultrasonic inspection* (*USI*): Inevitably, fatigue and other failure mechanisms initiate cracks in the rails. In order to detect such cracks, ultrasonic monitoring cars have been devised. USI are carried out to detect cracks before they develop to rail breakage.
- *Track circuit detection*: For some networks, track circuits have been implemented as a part of the signalling system. A convenient 'extra effect' of this system is that it allows detecting rail breakage. The track circuit is a part of the centralised traffic control (CTC) system. Only breakages that result in a complete failure of the rail may be detected by this system. The CTC system is useful to detect rail breakages, which USI may have missed.
- *Physical barriers*: If a breakage is not detected, derailment may occur. However, there are some physical barriers such as 'guide rails' used on bridges, in curves and tunnels, etc. which allow maintaining a veering train on the tracks.

This paper deals with the optimisation of the 'USI barrier'.

2.1. Influence diagram for ultrasonic inspections

The model describing the effects that USI have on the safety and on the economics of the rail system can be described with the aid of the influence diagram in Fig. 2. Influence diagrams [3] are excellent tools comprised of nodes and arrows to visualise the qualitative relationships existing among the variables of a system. More precisely, an influence diagram captures the relations between decisions (e.g. maintenance actions), random quantities (e.g. number of broken rails), and values (e.g. cost of derailments). Correspondingly, there are three types of nodes:

- *Decision nodes*, drawn as rectangles, represent states for which the decision maker is in a position to choose an action, e.g. the type and frequency of maintenance.
- *Chance nodes*, drawn as circles, represent random quantities and are influenced by the decision nodes.
- *Value nodes*, drawn as rounded rectangles, represent the outputs of the system, e.g. costs for maintenance or damages.

The arrows represent the qualitative relations (influences) between the nodes in the diagram, i.e. causes and effects. The diagram provides a basis for setting up the quantitative model.

The influence diagram for the USI problem is shown in Fig. 2. Table 1 explains the labels of all nodes used in the influence diagram of Fig. 2. The only decision node in the diagram is the 'inspection strategy' node in the top-left part of the diagram. The node influences directly the costs for USI and the number of detected cracks per year (no. det. cracks by USI). This number is also influenced by the length of the P–F-interval (see Section 2.2 for its definition) and by the effectiveness of the ultrasonic measuring technique, in terms of its ability to detect a possible defect.

Let us now follow the influence diagram implication relationships starting from the bottom-left chance node relative to the number of rail cracks per year (no. rail cracks), whose corresponding random variable we denote by f_I. The quantity f_I depends on operational and environmental factors such as the traffic load, the speed of the trains, the season, the quality of the materials, the superstructure quality and others. Empiric models are available to evaluate f_I as a function of the above factors [6]. In order to calibrate the parameters feeding such

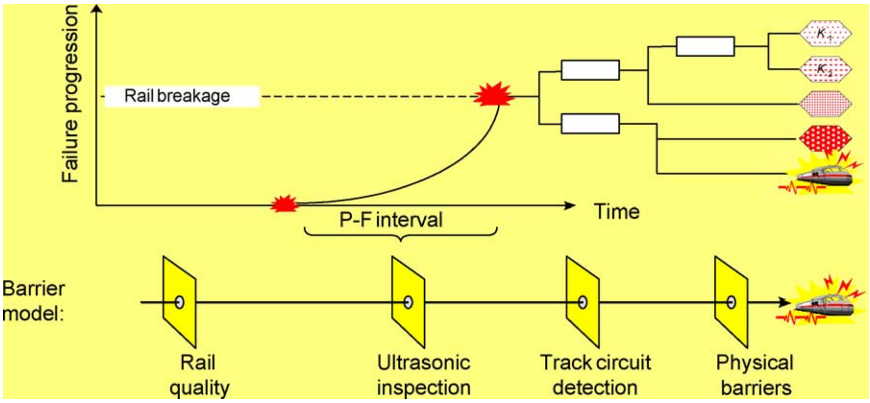


Fig. 1. The multiple barrier model.

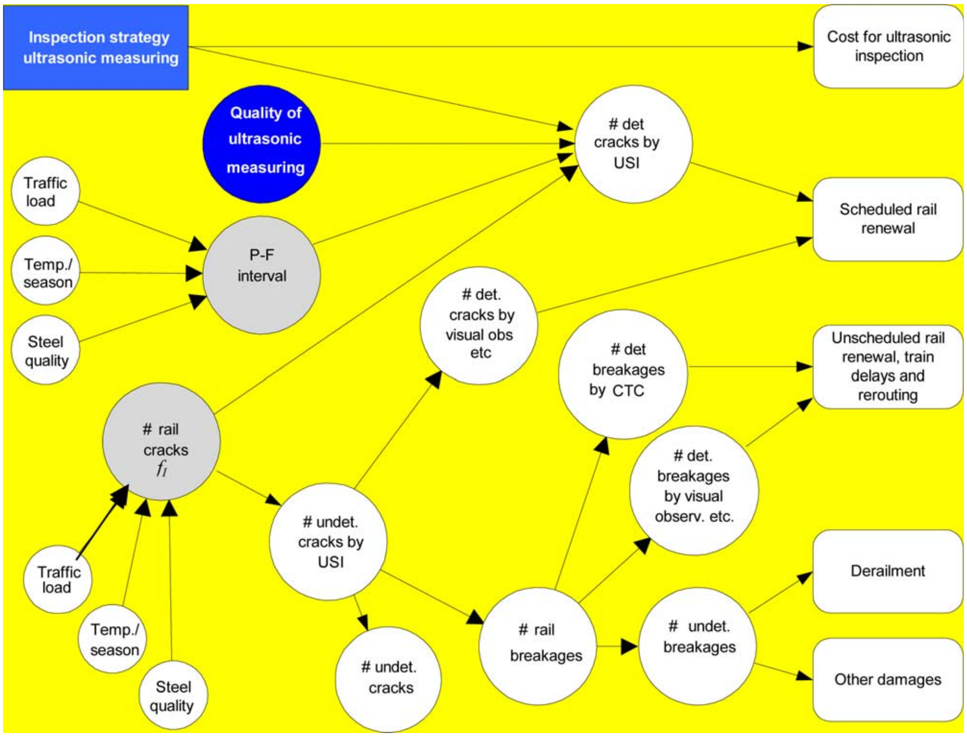


Fig. 2. Influence diagram for ultrasonic inspection.

models, a comprehensive data analysis is currently being performed by JBV, which specifically addresses the role of operational and environmental factors on the initiation and growth rates of tracks defects. Partial results from this study were used to assess a representative value of f_I for the present study. The above mentioned operational and environmental factors influence also the length of the P–F interval. For clarity of the graphical representation, in the influence diagram only a few of these influencing variables are shown.

Table 1
Labels used in the influence diagram

Label	Explanation
Inspection strategy ultrasonic measuring	Rail Inspection strategy. To be optimised in this work
Traffic load; Temperature/Season; Steel quality	Operational and environmental factors that influence the length of the P–F interval and the frequency of crack initiations
P–F interval	Length of the P–F interval
No. rail cracks, f_I	Number of new rail cracks per year
No. det. cracks by USI	Number of cracks per year detected by USI
No. undet. cracks by USI	Number of cracks per year not detected by USI
No. det. cracks by visual obs., etc.	Number of cracks per year detected by visual observation, by train personnel, during rail maintenance, and other
No. undet. cracks	Number of cracks per year, which remain undetected in the rail
No. rail breakages	Number of rail breakages per year
No. det. breakage by CTC	Number of breakages per year detected by Centralised Train Control (CTC) (track circuit detection)
No. det. breakage by visual observ., etc.	Number of breakages per year detected by visual observation, by train personnel, by USI, and during rail maintenance.
No. undet breakages	Number of undetected rail breakages per year.
Costs for ultrasonic inspection	Costs for ultrasonic inspection (measuring wagon and personnel)
Scheduled rail renewal	Costs for material and personnel
Unscheduled rail renewal, train delays and rerouting	Costs for material, personnel, and traffic restrictions.
Derailment	Costs for damages both to track and rolling stock, injured and killed persons.
Other damages	Costs for damages e.g. to wheels and other parts of the rolling stock.

Continuing the analysis of the influence diagram in Fig. 2, we see that some cracks are detected by USI (no. det. crack USI) while some others are not (no. undet. crack USI). However, some of these undetected cracks can be detected by other means, e.g. by visual inspection (no. det. cracks by visual obs., etc). A crack detected either by USI or by visual observation is subject to scheduled rail renewals. Not all of the undetected cracks lead to rail breakage (no. rail breakages). Others (no. undet. cracks), can remain undetected in the rail, though without developing into rail failure. A large portion of the rail breakages can be detected by Centralised Train Control (no. det. breakages by CTC) and other detection methods such as visual inspection by train personnel, etc. (no. det. breakages by visual observ., etc). The repair of a detected breakage constitutes an unplanned rail renewal, with additional expenses associated to train delays and rerouting. Finally, those rail breakages remaining undetected lead to derailments or other damages.

2.2. The P–F interval and the associated discrete-states crack model

The concept of P–F interval is often used in the analysis of maintenance procedures. An exhaustive review of models based on the P–F interval concept can be found in Ref. [5]. The P–F interval is defined as the interval between the time that a potential failure can be detected and that at which the failure occurs. In practical terms, then, the P–F interval is the grace time available to detect the presence of a defect before it causes the actual failure. The basic idea behind this concept is that of considering the failure as a two-stage process (Fig. 3, on the left): first, at some time a defect in the system becomes detectable; then, after some delay-time the system fails due to the degeneration of the defect.

More specifically, assume that a crack originates at time T_{init}. The defect may not be detectable from the time it is initiated, but only at time T_{det} when its progression reaches the point '*P*', i.e. the point in time when the defect has reached a dimension such that the measurement machine is able to sense it. If the failure progression is not stopped, it may reach at time T_{crit} some critical value, corresponding to which a breakage/failure occurs (point '*F*').

A common hypothesis assumed in the P–F interval modelling approach is that the probability of detection of a defect does not depend on the time since the defect has originated. This is an unrealistic hypothesis, which we shall drop in our model, since such probability is, in most cases, dependent on whether the defect is newly born or it is well developed [6]. Moreover, in our case of interest, the set up of the model for the rail failure probability is further complicated by the need of accounting for the actual inspection and maintenance procedures followed in the practice of railway operation. In particular, at JBV, the track condition is inspected periodically, τ being the inspection period, by running a wagon carrying the measuring instrumentation. Once a crack is detected it is assigned a 'severity class' to which there corresponds a particular maintenance procedure:

Failure class 2*b*:	Keep rail under observation
Failure class 2*a*:	Repair failure after a given waiting time t_{w}
Failure class *1*:	Repair failure quickly
Failure class 0:	Repair failure immediately and initiate traffic restrictions until failure is fixed

Whenever a crack is recognized to belong to class 2*b*, it is not repaired immediately, but simply periodically monitored by running a handled trolley every τ' units of time. If a crack is recognized to be of class 2*a*, its repair is delayed of a fixed time, t_{w}, usually dependent of the tonnage of the rail section. The economical advantage of delaying the repair action stands in the possibility of accumulating a certain number of cracks and opportunistically repair them all at once, thus reducing the economical burden of possible traffic restrictions or delays. Instead, for safety-related

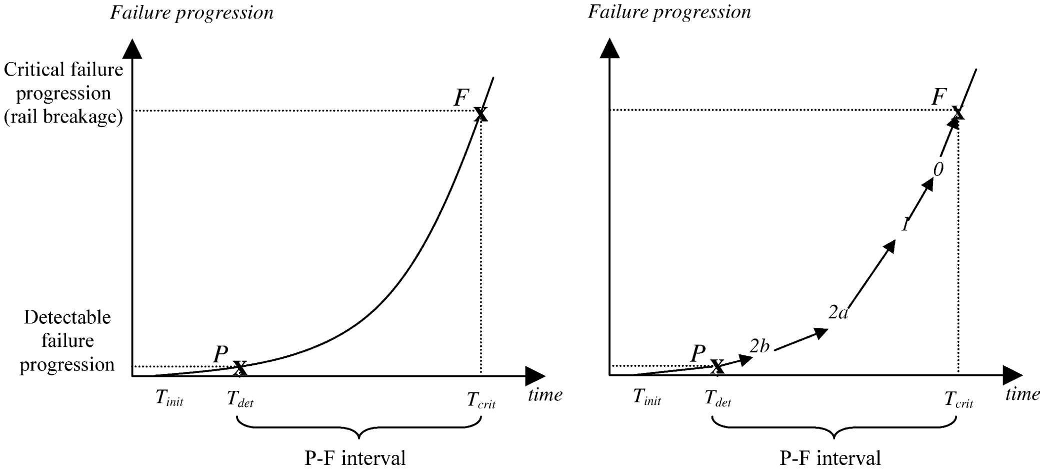


Fig. 3. Illustration of the P–F interval (left) and of the discrete-state model for cracks (right).

Table 2
Values of the transition rates λ and corresponding expected transition times $E[T]$

$\lambda_{2b\to 2a}$ ($E[T_{2b\to 2a}]$)	$\lambda_{2a\to 1}$ ($E[T_{2a\to 1}]$)	$\lambda_{1\to 0}$ ($E[T_{1\to 0}]$)	$\lambda_{0\to F}$ ($E[T_{0\to F}]$)
1 [y^{-1}] (1.00 [y])	1.47 [y^{-1}] (0.68 [y])	0.38 [y^{-1}] (2.60 [y])	1.33 [y^{-1}] (0.75 [y])

reasons, the repairs of cracks of severity class *1* or 0 cannot be delayed, though there is a difference in the priority of the two actions. For those of class *1* the repair is performed as soon as feasible but without any limitation to the train circulation in the corresponding section, while for those of class 0 the repair is categorically undertaken immediately, with traffic restrictions being imposed in the corresponding rail section until the crack is fixed.

A Markov model has been built to follow the crack evolution from T_{init} to T_{crit} through the different criticality conditions corresponding to the different failure severity classes considered at JBV (Fig. 3, on the right). According to the model, if not detected by USI, the crack propagates stochastically through the severity classes $2b$, $2a$, *1*, 0, until it possibly causes complete rail breakage (point *P*).

For simplicity, but without loss of generality, in our model we assume that the times T_{init}, T_{det} and the time at which the crack reaches the severity class $2b$ (Fig. 3) coincide. The values of the transition rates governing the Markov process have been taken such to adhere as much as possible to the information available at JBV (Table 2). More precisely, from the data available at JBV, the average time for a crack to propagate from failure class $2b$ to failure class 1 is expected to be 1.68 y with a standard deviation of 1.71 y and the average P–F interval T_{PF} is thought to be 5 y, with a standard deviation of 3 y. Correspondingly, the first piece of information is used to set the values for the transition rates $\lambda_{2b\to 2a}$ and $\lambda_{2a\to 1}$ in such a way that the average time for the crack to transfer from state $2b$ to state 1 is 1.68 y. As for the remaining values of the transition rates, reasonable values have been set such that the mean and standard deviation of the distribution of the random variable $T_{\text{PF}} = T_{2b\to 2a} + T_{2a\to 1} + T_{1\to 0} + T_{0\to F}$ are approximately 5 and 3 y, respectively.

2.3. Risk and cost assessment

For the quantitative evaluation of the economical and safety aspects involved in the decision process it is necessary to assess values for:

- the yearly frequency of crack initiations f_{I}
- the probability Q of USI failing to detect a crack, given that the crack has originated. The number of crack missed by USI per year is thus $f_{\text{I}}\,Q$ and of those detected is f_{I} (1-Q)
- the conditional probabilities which weigh the arrows in the diagram of Fig. 2
- the cost figures related to inspection, maintenance, damages and accidents, etc.

A comprehensive data analysis study is currently being undertaken by JBV to characterize quantitatively the degradation- and economics-related parameters feeding the model presented in this paper. For this reason, the level of detail of the model has been chosen accordingly with the foreseeable information from the JBV study. At present time, the main data sources available for assessing the values of the other parameters are:

- Norwegian railway statistics
- Experts from JBV (the Norwegian National Rail Administration)
- Others, e.g. literature, etc.

The frequency of cracks initiation f_{I} has been set to 30, which is regarded to be a representative value for a section of approximately 500 km. The values of the conditional probabilities for weighing the arrows in the diagram and of the cost figures are reported in Tables 3 and 4, respectively. We refer to [7] for a detailed description of the rationale of the procedure of assessing the values of the parameters involved in the optimisation model.

A particular USI strategy is identified by the triplet τ, τ', t_{w}. As mentioned before, the quantitative risk measure we adopt is the probability of USI failure to detect the crack, $Q(\tau,\ \tau',\ t_{\text{w}})$, which is directly related to the expected number of derailments per year, $E[D(\tau,\ \tau',\ t_{\text{w}})]$. Indeed, according to the implications stream identified by

Table 3
Additional parameter values used in the optimisation model

Parameter	Corresponding conditional probability
No. undet. cracks by USI → no. det. cracks by visual obs., etc.	0.1
No. undet. cracks by USI→ no. undet. Cracks	0.1
No. undet. cracks by USI → no. rail breakages	0.8
Sum:	1
No. rail breakages→no. det. breakages by CTC	0.4
No. rail breakages →no. det. breakages by visual observ., etc.	0.55
No. rail breakages→no. undet. Breakages	0.05
Sum:	1
No. undet. breakages→no. derailments	0.05
No. undet. breakages→no. other damages	0.95
Sum:	1

Table 4
Costs per event of USI, rail renewal, derailment and other damages

Value node	Basic costs
Cost for ultrasonic inspection (train)	200,000 NOK
Cost for ultrasonic inspection (handheld trolley)	5000 NOK
Scheduled rail renewal	15,000 NOK
Rail renewal, when postponed of time t_w	15,000/ n_w NOK
Unscheduled rail renewal, train delays, and rerouting	40,000 NOK
Derailment (includes possible human losses)	15,000,000 NOK
Other damages	40,000 NOK

1 € ≈ 8 NOK.

the influence diagram of Fig. 2:

$$
\begin{aligned}
E[D(\tau,\tau',t_w)] = {} & \\
& f_I Q(\tau,\tau',t_w)\cdot \Pr[\text{rail breakage}\,|\text{crack undetected by USI}] \\
& \cdot \Pr[\text{undetected breakage}|\text{rail breakage}] \\
& \cdot \Pr[\text{derailment}|\text{undetected breakage}] \qquad (1)
\end{aligned}
$$

In view of the common definition of risk as a set of triplets {scenario, probability, consequence}, the scenarios we consider in the present paper are those which lead to train derailment, which is taken as the consequence. Other consequences may be considered, ranging from minor damage to human losses. In such cases the influence diagram of Fig. 2 should be modified accordingly.

The economical impact of a particular inspection strategy is evaluated considering the following yearly costs: USI performed both by the measuring wagon and by the manual trolley; scheduled, unscheduled and postponed rail renewals; derailment and other damages. Table 4 reports the costs associated to the various events as provided by JBV and other US data [7]. Note that, as mentioned earlier, if the repair of a crack is postponed of a time t_w, then the costs of the renewal action relative to that crack are lower, since other cracks may have been detected during t_w and the repair costs can be shared among all of them. To asses a cost figure representative of this situation we consider the expected number n_w of crack initiations during time t_w. The quantity n_w can be assessed as $n_w = f_I t_w$ and the cost of the renewal action per crack is the cost of rail renewal divided by n_w and the overall expected yearly cost associated to a given USI strategy is calculated in a similar way as in Eq. (1).

Only point estimates are considered in this paper for the parameter values, thus neglecting the effects of uncertainties on the optimization and decision-making processes. Note that the model is suitable for uncertainty propagation by Monte Carlo sampling procedures.

3. The model for periodic inspections

In this Section we describe a general method for including inspections into Markov models [9]. In the next Section this modelling framework is tailored to the present case study. Semi-Markov models have also been used to describe systems undergoing periodic test and maintenance. For example, in Ref. [26] the case of systems undergoing specific changes of state at predetermined instances of time and transiting to states with generally distributed times is considered.

When the system is subject to inspections at predetermined times, two types of transitions may occur among the m states of the system (state 1 denoting the nominal state). Transitions of the first type occur randomly and, if the system is Markovian, have exponentially distributed transition times. Such transitions are characterized by constant transition rates $\lambda_{i\to j}$, $i,j = 1,2,\ldots,m$ which constitute the transition matrix Λ. Transitions of the second type account for the state changes occurring at the times of inspection, t_r, $r = 1,2,\ldots$ and thus are characterized by transition rates of the form $l_{i\to j}\cdot\delta(t\text{-}t_r)$ which constitute the characteristic inspection matrix $\Gamma\delta(t-t_r)$, where the elements of Γ specify the effects of the inspections and maintenances on the system. As illustrated in the following Section, imperfect inspections and different maintenance strategies can be included in this modelling scheme. The set of equations governing the evolution of the system state probability vector $\underline{p}(t) = [p_1(t), p_2(t), \ldots, p_m(t)]$ is:

$$
\begin{aligned}
& \frac{\mathrm{d}\underline{p}(t)}{\mathrm{d}t} = \underline{p}(t)\Lambda + \underline{p}(t)\Gamma\delta(t-t_r) = \underline{p}(t)M \quad r = 1,2,\ldots \\
& \underline{p}(0) = [1,0,\ldots,0]
\end{aligned} \qquad (2)
$$

where $M = \Lambda + \Gamma\delta(t\text{-}t_r)$ is the total system transition matrix. Eq. (2) implies for the jth component of the state probability vector:

$$
\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} = \sum_{i=1}^{m} \lambda_{i\to j} p_i(t) + \sum_{i=1}^{m} l_{i\to j} p_i(t)\delta(t-t_r) \quad r = 1,2,\ldots \qquad (3)
$$

The first term of the above Eq. (3) accounts for changes in $p_j(t)$ due to stochastic transitions occurring at exponentially distributed transition times. The second term accounts for changes in $p_j(t)$ due to instantaneous transitions occurring only at the times of the inspection.

Closed form solutions to Eq. (3) can be derived for quantities such as the mean time to failure and the mean availability [9]. These analytical expressions are based on inverse and exponential matrices and become difficult to handle when the complexity of the system increases. In these latter cases, one can resort to standard finite-difference approximations.

Table 5
Probabilities of independent and CCF of USI for different states of the crack

State (i)	$2b$	$2a$	1	0
$q_I(i)$	0.35	0.15	0.05	0.02
$q_C(i)$	0.15	0.10	0.08	0.05

4. Estimating the rail breakage probability

In this Section we specify the model developed in this work for estimating the probability of USI failure to detect a crack, $Q(\tau, \tau', t_w)$. The model, tailored on the modelling framework described in the previous Section 3, gives credit to:

- the different maintenance procedures that are followed in correspondence with each failure class considered at JBV
- dependence of the probability of not detecting a crack on its degradation condition
- CCF due to systematic errors in the inspection procedure.

As explained earlier, the physical basis of the model is the progression of the crack through the severity class $2b$, $2a$, 1, 0, captured by the Markov model on the right of Fig. 1. If the crack remains undetected during its progression, it may develop to rail breakage (condition F). On the contrary, if it is revealed by USI, a specific maintenance action is undertaken depending on the identified criticality condition.

Every τ time units the rail is inspected and, if the inspection is successful, the crack is revealed. Inspection can fail due to:

- conditions specific to the particular inspection run (independent failures), e.g. inadequate velocity of the measuring wagon, human errors
- systematic conditions which prevent the detection of the crack in any run (Common Cause Failures, CCF), e.g. low coverage

For USI, the probabilities of independent and common cause failures are $q_I(i)$ and $q_C(i)$, $i=2b$, $2a$, 1, 0, respectively. Reasonable values for $q_I(i)$ and $q_C(i)$, are shown in Table 5 (adapted from [6]). Note that the failures probabilities decrease with the increase of the crack severity condition: realistically, the detection of a newly born crack, corresponding to a low severity condition, can be trickier than that of an already developed crack.

For clarity purposes, we consider at first only independent failures in the description of the model [8]. Then, we illustrate how to incorporate CCF. The state transition diagram, without common cause failures, is reported in Fig. 4. To distinguish the deterministic system transitions resulting from the rail inspections and repairs from the usual stochastic transitions, the former are represented by dashed lines whereas the latter are represented by solid lines. The number of states needed for the description of the entire process is $m=18$. The states numbering and the notation used is explained in Table 6. States describing CCF of USI are also reported in the Table, even though their role will be addressed later on.

4.1. The physical model

Three main paths can be identified in the state transition diagram of Fig. 4. The first one is the 'undetected' path $2b \rightarrow 2a \rightarrow 1 \rightarrow 0 \rightarrow F$, identified by the first five states listed in Table 6: if not detected by any inspection, the crack visits all of the severity classes from $2b$ to 0 before possibly

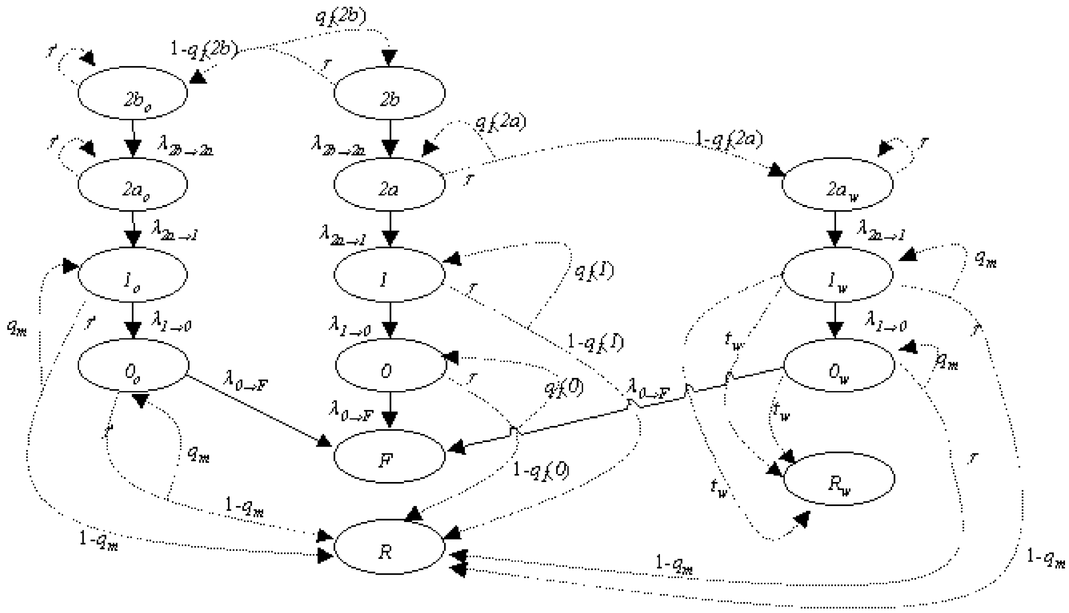


Fig. 4. State transition diagram (no common cause failures).

Table 6
Numbering of the Markov states for the complete model

State no.	Description	Nomenclature
1	Failure class 2*b*	$2b$
2	Failure class 2*a*	$2a$
3	Failure class 1	1
4	Failure class 0	0
5	Failure	F
6	Crack revealed and repaired	R
7	Failure class 2*b*, under observation	$2b_{\mathrm{o}}$
8	Failure class 2*a*, under observation	$2a_{\mathrm{o}}$
9	Failure class 1, under observation	1_{o}
10	Failure class 0, under observation	0_{o}
11	Failure class 2*a*, waiting for repair	$2a_w$
12	Failure class 1, waiting for repair	1_w
13	Failure class 0, waiting for repair	0_w
14	Crack revealed and repaired, waiting	R_w
15	Failure class 2*b*, common cause failure	$2b_c$
16	Failure class 2a, common cause failure	$2a_c$
17	Failure class 1, common cause failure	1_c
18	Failure class 0, common cause failure	0_c

developing to rail breakage (condition F). The second path is the 'observed' one $2b_{\mathrm{o}}\rightarrow 2a_{\mathrm{o}}\rightarrow 1_{\mathrm{o}}\rightarrow 0_{\mathrm{o}}\rightarrow F$, identified by the states 7–10 in Table 6 and describes the crack evolution when a failure class $2b$ is detected and the crack is subject to the corresponding maintenance policy, i.e. it is kept under observation and inspected every τ' time units by means of a manual trolley. The crack then can proceed its degradation process until it is repaired when recognized to be either in severity class 1 or 0. Upon repair, the crack transfers to state 6 (condition R): the potential failure, detected when the crack severity class was $2b$, has been avoided. On the contrary, rail breakage (condition F) can occur if the crack manages to propagate through the whole degradation path $2b_{\mathrm{o}}\rightarrow 2a_{\mathrm{o}}\rightarrow 1_{\mathrm{o}}\rightarrow 0_{\mathrm{o}}\rightarrow F$. This situation occurs if the crack succeeds in propagating from either of the severity classes $2b_{\mathrm{o}}$ or $2a_{\mathrm{o}}$, which require no immediate repair, to the condition F within the time interval τ'. Alternatively, rail breakage can occur due to errors in the classification of the crack severity during inspections with the manual trolley: a crack in severity class 1_{o} or 0_{o} may not be recognized as such so that the necessary immediate repair action is not performed. The probability of misinterpreting the crack condition is here taken to be $q_{\mathrm{m}}=10^{-3}$, following the suggestion of experts at JBV. The third path running along the states 11–14 ($2a_{\mathrm{w}}$, 1_{w}, 0_{w}, F, R_{w}) characterizes the 'on wait' crack evolution when the revealed crack severity class is classified as $2a$. If this is the case, according to JBV maintenance procedures, the crack is repaired to state R_{w} after a time t_{w} and the rail breakage is avoided. However, rail breakage can occur if the crack propagates form $2a_{\mathrm{w}}$ to F within $t_{\mathrm{w}.}$ Including state R_{w} is necessary to evaluate the probability of repairing an 'on wait' crack, to which a specific renewal cost figure is associated (Section 2). In any case, if the successive inspection is performed before t_{w} and the severity classes 1_{w} or 0_{w} are revealed, the crack is repaired immediately at the time of the successive inspection. Finally, if the USI detects a crack in severity class 0 or 1, it is repaired immediately, i.e. an instantaneous transition occurs to state R.

4.2. Algorithm description

The proposed model stands on the four transition matrices, Λ, L, L', and L_{w}. From the point of view of the algorithm, at $t\neq t_k,\ t\neq t_l,\ t\neq t_k+t_w,\ t_k=k\tau+t_0,\ k=1,2,\ldots,\ t_l=l\tau'+t_0,\ l=1,2,\ldots$, where t_0 is the time of the first inspection, the crack evolution is governed by the time-independent transition rate matrix Λ. Such matrix describes the degradation process through the three different paths $2b\rightarrow 2a\rightarrow 1\rightarrow 0\rightarrow F$, $2b_{\mathrm{o}}\rightarrow 2a_{\mathrm{o}}\rightarrow 1_{\mathrm{o}}\rightarrow 0_{\mathrm{o}}\rightarrow F$ and $2a_{\mathrm{w}}\rightarrow 1_{\mathrm{w}}\rightarrow 0_{\mathrm{w}}\rightarrow F$ and is built upon the transition rates $\lambda_{i\rightarrow j}$, $i, j=2b,\ 2a,\ 1,\ 0$. On the other hand, deterministic transitions occur at times $t=t_k,\ t=t_l,\ t=t_k+t_{\mathrm{w}}$, and the inspection-transition matrices $\Gamma\equiv L$ or L' or L_{w} are correspondingly applied to the state probability vector p to specify the effects of the particular action of inspection/maintenance on the crack state. The elements of these matrices are the conditional probabilities $q_{\mathrm{I}}(i)$, $i=2b,\ 2a,\ 1,\ 0$ and the probability of misinterpreting the criticality condition of a crack, q_{m}. The former probabilities govern the instantaneous 'jumps' out of the main path $2b\rightarrow 2a\rightarrow 1\rightarrow 0\rightarrow F$ following the inspection while the latter governs 'jumps' from the two secondary paths $2b_{\mathrm{o}}\rightarrow 2a_{\mathrm{o}}\rightarrow 1_{\mathrm{o}}\rightarrow 0_{\mathrm{o}}\rightarrow F$ and $2a_{\mathrm{w}}\rightarrow 1_{\mathrm{w}}\rightarrow 0_{\mathrm{w}}\rightarrow F$ to the repaired state R_{w}.

The recursive Eq. (2) is then used to compute the time evolution of the state probability vector $p(t)$ for each value of the discretized time variable $t=0,\ldots,\tau$. For example, at the time of periodic inspection, $t_k=k\cdot\tau+t_0,\ k=1,2,\ldots$, the state probability vector p is multiplied by matrix $\Gamma=L$ so that the probabilities $p_j(t)$ of being in state j at time $t=t_k$ are modified to account for possible inspection strategies. For instance, considering the possible instantaneous transfer from state $2b$ to state $2b_{\mathrm{o}}$, sketched by the dotted arrows in the top-left part of Fig. 4, the multiplication implies for the probabilities of the states $2b$ and $2b_{\mathrm{o}}$:

$$p_{2b}(t_k)\rightarrow p_{2b}(t_k)q_I(2b) \tag{4}$$

$$p_{2b_o}(t_k)\rightarrow p_{2b_o}(t_k)+p_{2b}(t_k)(1-q_I(2b)) \tag{5}$$

Eq. (4) describes the fact that the crack in state 1 (severity class $2b$) at a time t_k corresponding to an inspection, has a probability $q_{\mathrm{I}}(2b)$ of not being revealed and, correspondingly, of remaining in the same state 1, so that the probability of residing in such state, $p_{2b}(t_k)$ is switched to $p_{2b}(t_k)q_{\mathrm{I}}(2b)$. In this case, from time t_k to the successive inspection time $t_{k+1}=t_k+t$, the crack continues to propagate along the path $2b\rightarrow 2a\rightarrow 1\rightarrow 0\rightarrow F$, with stochastic transitions governed by the matrix Λ. Correspondingly, Eq. (5) shows that, while in severity class $2b$, the crack has a probability $1-q_{\mathrm{I}}(2b)$ of being revealed by the USI, i.e. of instantaneously transferring to severity class $2b_{\mathrm{o}}$ (state 7) and then propagating, 'under observation' along $2b_{\mathrm{o}} \rightarrow 2a_{\mathrm{o}} \rightarrow 1_{\mathrm{o}} \rightarrow 0_{\mathrm{o}} \rightarrow F$. The other changes in the state probability values needed to reproduce the other instantaneous transitions occurring at the inspection time t_k, (dotted lines labelled by 'τ' in Fig. 4) are obtained likewise by multiplying the state probability vector p times the properly shaped matrix $\Gamma = L$.

When a crack is recognized to be in severity class $2b$, it is kept under observation and it is inspected every τ' time units (Fig. 4). Then, the crack is repaired when it is recognized to be in severity class $2a$ or higher. To account for these additional periodic inspections, we use the inspection-transition matrix $L'\delta(t - t_l)$, $t_l = lt'$, $l = 1, 2, \ldots$ When these inspections are performed, there is a probability q_{m} of misinterpreting the crack condition. We assume that q_{m} is the probability of not recognizing that the crack is in either of the states 1_{o} or 0_{o} which require a repair action. At time $t = t_l$, $l = 1, 2, \ldots$, the inspection-transition matrix L' modifies the following state probabilities $p_i(t)$ pertaining to states within the path $2b_{\mathrm{o}} \rightarrow 2a_{\mathrm{o}} \rightarrow 1_{\mathrm{o}} \rightarrow 0_{\mathrm{o}} \rightarrow F$:

$$\begin{aligned} p_{1_o}(t_l) &= p_{1_o}(t_l) q_{\mathrm{m}} \\ p_{0_o}(t_l) &= p_{0_o}(t_l) q_{\mathrm{m}} \quad l = 1, 2, \ldots \end{aligned} \tag{6}$$

When a criticality severity class $2a$ is revealed, the crack stochastically runs through the path $2a_{\mathrm{w}} \rightarrow 1_{\mathrm{w}} \rightarrow 0_{\mathrm{w}} \rightarrow F$ for a waiting time t_{w} after which a repair action is performed to bring the crack to state 14 (R_{w}). This is done by applying to p the inspection-matrix L_{w} at times $t = k\tau + t_{\mathrm{w}}$ in a way similar to those already explained for the matrices L and L'.

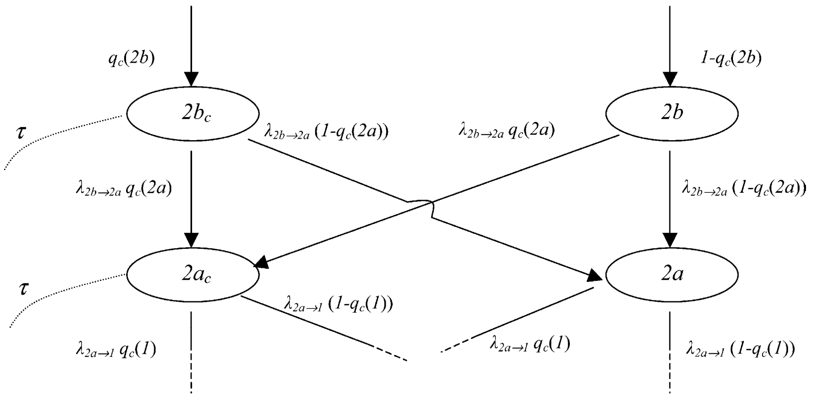
4.3. Incorporating common cause failures

As previously mentioned, besides the independent failure modelled above, the USI can fail also due to common cause failures, for example originated by systematic errors in the inspection procedures. In such cases, the effects of a USI common cause failure is not confined to a particular inspection run of the measuring wagon but it affects all of the runs. In other words, due to a USI common cause failure, the potential rail breakage cannot be revealed in any inspection runs. Furthermore, the probability of CCF, $q_{\mathrm{C}}(i)$, $i = 2b, 2a, 1, 0$, is dependent on the current severity condition i of the crack. In our modelling framework, the USI common cause failures are described by the last four states 15–18 (severity classes $2b_{\mathrm{c}}$, $2a_{\mathrm{c}}$, 1_{c}, 0_{c}) and the modelling of the crack progression along the main path $2b \rightarrow 2a \rightarrow 1 \rightarrow 0 \rightarrow F$ of Fig. 4 is modified as in Fig. 5. The initial state $2b$ is now mirrored by state $2b_{\mathrm{C}}$. A newly born crack can be exclusively in either one of the states $2b$ or $2b_{\mathrm{C}}$ with probability $1 - q_{\mathrm{C}}(2b)$ and $q_{\mathrm{C}}(2b)$, respectively. If the crack is in state $2b$, as before, at the time of the inspection it has a probability $1 - q_{\mathrm{I}}(2b)$ of being revealed and, accordingly, of transferring to the path $2b_{\mathrm{o}} \rightarrow 2a_{\mathrm{o}} \rightarrow 1_{\mathrm{o}} \rightarrow 0_{\mathrm{o}} \rightarrow F$. On the contrary, as long as the crack is in state $2b_C$ the inspections are ineffective. If the crack is not revealed while in severity class $2b$, it will eventually worsen to severity class $2a$. For a crack in severity class $2a$, there is a probability $q_{\mathrm{C}}(2a)$ that all inspections fail due to a CCF. Accordingly, the rates of transitions from the states $2b$ and $2b_{\mathrm{C}}$ to the state $2a_{\mathrm{C}}$ are both equal to $\lambda_{2b \rightarrow 2a} q_{\mathrm{I}}(2a)$ and those to the state $2a$ are both equal to $\lambda_{2b \rightarrow 2a}(1 - q_{\mathrm{I}}(2a))$. Again, the USI can reveal the crack if it is in state $2a$, whereas it is ineffective as long as the crack resides in state $2a_{\mathrm{C}}$. The successive evolution of the crack through states 1, 0, F, 1_{C}, 0_{C} and F_{C} then proceeds in a similar manner.

4.4. Specific comments

Note that the relative position of the time T_{init} at which the crack appears within the generic inspection period of duration τ is not known and the probability of revealing a crack is indeed different if the crack has initiated just before of just after the generic inspection. To approximate this effect, we consider ten different possible times of first inspection t_0, equidistant on the interval $(0,\tau)$, and compute the corresponding values of USI failure probability. Then, we take the average as the representative value for Q.

Furthermore, Note that the model thereby obtained is time-inhomogeneous, i.e. the transition probability from a state at time t to another time t' does not depend only on



$q_c(2b)$
$1\text{-}q_c(2b)$
$2b_c$
$2b$
τ
$\lambda_{2b\rightarrow 2a}\,(1\text{-}q_c(2a))$
$\lambda_{2b\rightarrow 2a}\,q_c(2a)$
$\lambda_{2b\rightarrow 2a}\,q_c(2a)$
$\lambda_{2b\rightarrow 2a}\,(1\text{-}q_c(2a))$
τ
$2a_c$
$2a$
$\lambda_{2a\rightarrow 1}\,(1\text{-}q_c(1))$
$\lambda_{2a\rightarrow 1}\,q_c(1)$
$\lambda_{2a\rightarrow 1}\,q_c(1)$
$\lambda_{2a\rightarrow 1}\,(1\text{-}q_c(1))$

Fig. 5. Modification of the state transition diagram for the USI common cause failure.

the difference $t'-t$. Indeed, consider for example the probability density $P(2b\rightarrow 2a,\ t,\ t')$ that a crack in severity class $2b$ at time t transfers to severity class $2a$ at time t', s.t. $t'-t'\tau$, for simplicity without considering the effect of CCF:

$$P(2b\rightarrow 2a,t,t') = \exp(-\lambda_{2a\rightarrow 2b}(t'-t)) \qquad \text{if } t_{k\notin}[t,t']$$

$$P(2b\rightarrow 2a,t,t') = q(2a)\exp(-\lambda_{2a\rightarrow 2b}(t'-t)) \quad \text{if } t_{k\in}[t,t'] \tag{7}$$

From Eq. (7), $P(2\text{b}\rightarrow 2\text{a},\ t,\ t')$ depends on both the difference $t'-t$ and on whether the generic inspection at time t_k occurs within the interval $[t,t']$.

5. The multi-objective genetic algorithm optimization approach

5.1. Genetic algorithms

In the following Sections we provide some basics of the GA optimization approach with reference to the traditional single-objective GA [12] and then present the generalization of the approach to multiobjective problems [23,24]. For a more detailed description of GAs, we refer to the extensive literature available on the subject.

5.1.1. Single objective genetic algorithms

Genetic algorithms, first formalized as an optimization method by Holland [11], are search tools modeled after the genetic evolution of natural species. The GAs differ from most optimization techniques because of their global searching effectuated by one population of solutions rather than from one single solution. Every proposal of solution is represented by a vector $\underset{\sim}{X}$ of the independent variables, which is coded in a so-called chromosome, constituted by so-called genes, each one coding one component of $\underset{\sim}{X}$; a binary coding is widely used.

The GA search starts with the creation of a random initial population of N_{ga} chromosomes, i.e. potential solutions to the problem. Then, these individuals are evaluated in terms of their so-called fitnesses, i.e. of their corresponding objective function values. This initial population is allowed to evolve in successive generations through the following steps:

(1) selection of a pair of individuals as parents;
(2) crossover of the parents, with generation of two children;
(3) replacement in the population, so as to maintain the population number N_{ga} constant;
(4) genetic mutation.

Every time a new solution $\underset{\sim}{X}$ is proposed by the GA the objective function is evaluated and a ranking of the individuals in the current population is dynamically updated, based on their fitness values. This ranking is used in the selection procedure which is performed in such a way that in the long run the best individuals will have a greater probability to be selected as parents, in resemblance to the natural principles of the 'survival of the fittest'. Similarly, the ranking is used in the replacement procedures to decide who, among the parents and the daughters, should survive in the next population. An algorithm based on these procedures is often referred to as a steady-state genetic algorithm [18].

When using GAs, sufficient genetic diversity among solutions in the population should be guaranteed. Lack of such diversity would lead to a reduction of the search space spanned by the GA and consequently to a degradation of its optimization performance with selection of mediocre individuals resulting in premature convergence to a local minimum. On the other hand, an excess of genetic diversity, especially at later generations, may lead to a degradation of the optimization performance, resulting in very late, or even no, convergence.

5.1.2. Multiobjective genetic algorithms

The multiobjective optimization problem arises when in correspondence of each point $\underset{\sim}{X}$ in the search space we must consider several objective functions $f_i(\underset{\sim}{X})$, $i=1,2,\ldots,N_{\text{f}}$, and then identify that $\underset{\sim}{X}^*$ which gives rise to the best compromise among the various objective functions. The comparison of two solutions with respect to several objectives may be achieved through the introduction of the concepts of *Pareto optimality* and *dominance* [23,24] which enable solutions to be compared and ranked without imposing any a priori measure as to the relative importance of individual objectives, neither in the form of subjective weights nor arbitrary constraints.

Let us consider N_{f} different objective functions $f_i(\underset{\sim}{X})$, $i=1,2,\ldots,N_f$ where $\underset{\sim}{X}$ represents the vector of independent variables identifying a generic proposal of solution. We say that solution $\underset{\sim}{X}$ *dominates* solution $\underset{\sim}{Y}$ if $\underset{\sim}{X}$ is better on all objectives [20], i.e. if

$$f_i(\underset{\sim}{X}) > f_i(\underset{\sim}{Y}) \quad \text{for } i=1,\ldots,N_f$$

The solutions not dominated by any other are said to be *non-dominated* solutions.

Within the genetic approach, in order to treat simultaneously several objective functions, it is necessary to generalize the single-fitness procedures employed in the single-objective GA by assigning N_{f} fitnesses to each $\underset{\sim}{X}$.

Concerning the insertion of an individual (i.e. an $\underset{\sim}{X}$ value) in the population, often constraints exist which impose restrictions that the candidate individual has to satisfy and whose introduction speeds up the convergence of the algorithm, due to a reduction in the search space. Such constraints may be handled, just as in the case of single-objective GAs, by testing whether, in the course of the population creation and replacement procedures, the candidate solution fulfills the criteria pertaining to all the N_{f} fitnesses.

Once a population of individuals (chromosomes) $\{\underset{\sim}{X}\}$ has been created, we rank them according to the Pareto dominance criterion by looking at the N_f-dimensional space of the fitnesses $f_i(\underset{\sim}{X})$, $i=1,2,\ldots,N_f$, (see Fig. 6 for $N_f=2$). All nondominated individuals in the current population are identified. These solutions are considered the best ones, and assigned the rank 1. Then, they are virtually removed from the population and the next set of non-dominated individuals are identified and assigned rank 2. This process continues until every solution in the population has been ranked.

The selection and replacement procedures of the multi-objective genetic algorithms are based on this ranking: every chromosome belonging to the same rank class has to be considered equivalent to any other of the class, i.e. it has the same probability of the others to be selected as a parent and survive the replacement.

During the optimization search, an archive of vectors, each one constituted by a non-dominated chromosome and by the corresponding N_f fitnesses, representing the dynamic Pareto optimality surface is recorded and updated. At the end of each generation, nondominated individuals in the current population are compared, in terms of the fitnesses, with those already stored in the archive and the following archival rules are implemented:

1. If the new individual dominates existing members of the archive, these are removed and the new one is added;
2. if the new individual is dominated by any member of the archive, it is not stored;
3. if the new individual neither dominates nor is dominated by any member of the archive then:
   - if the archive is not full, the new individual is stored.
   - if the archive is full, the new individual replaces the *most similar* one in the archive. To do this, an appropriate concept of distance must be introduced to measure the similarity between two individuals: in this paper we shall adopt a euclidean distance based on the values of the fitnesses of the chromosomes normalized to the respective mean values in the archive.

The archive of nondominated individuals is also exploited by introducing an elitist parents' selection procedure which should in principle be more efficient. Every individual in the archive (or, if the archive's size is too large, a number of individuals given by a pre-established fraction of the population size N_{ga}, typically $N_{ga}/4$) is chosen once as a parent in each generation. This should guarantee a better propagation of the genetic code of non-dominated solutions, and thus a more efficient evolution of the population towards Pareto optimality, still preserving the genetic diversity.

At the end of the search procedure the result of the optimization is constituted by the archive itself, which gives the Pareto optimality region.

6. Optimisation results

The multi-objective optimization problem of our interest is to minimize the expected cost of yearly rail operation, C, and the rail probability of failure due to cracks, Q. The control (or decision) variables of the optimisation are:

- the inspection interval τ for USI performed by the main measurement wagon
- the inspection interval τ' for USI performed by the handled inspection trolley to monitor a crack detected in severity class $2b$
- the waiting time t_w before a repair action is performed when a crack is detected as in severty class $2a$.

Table 7
Genetic algorithm parameters and rules

Number of chromosomes in the population (population size)	75
Number of generations (termination criterion)	200
Selection	Fit–fit
Replacement	Children–parents
Mutation probability	0.001
Crossover probability	1
Number of generations without elitist selection	50
Fraction of parents chosen with elitist selection	0.25

In the framework of the multi-objective genetic algorithm procedure described in the previous Section 5, the generic potential solution $\underset{\sim}{X}$ is represented by a three-genes chromosome, coding the triplet (τ, τ' and t_w). As suggested by the experts, the range of the search is [4,20] months for all three variables, coded into 5-bit chromosomes. This entails a search space constituted by $32^3=32768$ candidate solutions. The data relevant for the multi-objective genetic algorithm procedure are contained in Table 7. The inverses of the yearly rail operation, C, and of the rail probability of failure due to cracks, Q, are the objective functions (fitnesses) to be maximized. The optimal solutions give rise to a two-dimensional Pareto set, non-dominated with respect to both cost and failure probability objectives.

Fig. 7 reports the values of the costs, C, and breakage probability, Q, in correspondence with all the 94

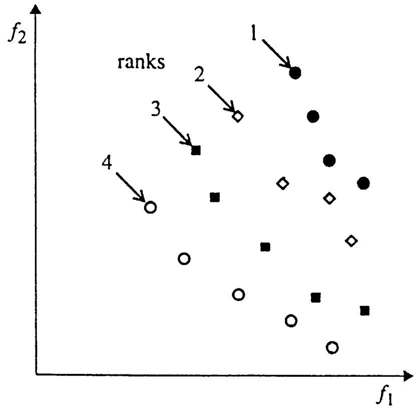


Fig. 6. Example of population ranking for a problem of maximization of f_1 and f_2.

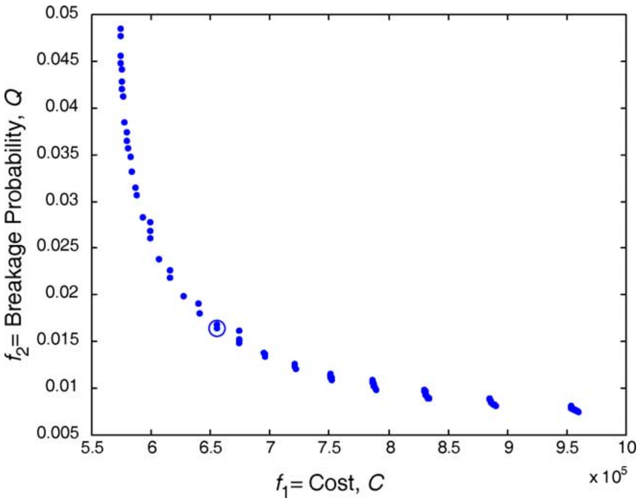


Fig. 7. Pareto front in the two-dimensional fitness space (C, Q). The symbol 'o' indicates the best compromise solution according to the min–max method.

non-dominated solutions contained in the archive at convergence. The CPU time required by the search was 5 min on an Athlon 1400 MHz computer. For validation purposes, the search space has also been crudely spanned and the solutions found by the genetic algorithm were verified to lie on the Pareto optimal boundary of the search space. The crude search took 1 h on the same machine. The significant difference in CPU times required by the crude search and the GA underlines the efficiency of the latter in driving quickly the search towards the portion of interest of the search space.

It can be seen in Fig. 7 that the non-dominated solutions tend to cluster so that the Pareto optimal frontier appears discontinuous. In order to get insights into this apparently anomalous behaviour, the non-dominated solutions can be sorted by increasing cost, i.e. by decreasing failure probability. The values of cost and failure probability corresponding to the non-dominated solutions ranked by increasing cost from 1 to 10, from 50 to 60 and from 84 to 94 are reported in Table 8. The first group of non-dominated solutions contains the best ten solutions with respect to the cost. The last group contains the best ten solutions with respect to the rail breakage probability. The middle group contains solutions representing a trade-off between low costs and low rail breakage probability. Table 8 also shows how the values of the control variables are arranged in this sorting of

Table 8
Values of cost and rail breakage probability of some non-dominated solutions found by the algorithm, sorted by increasing associated cost

Solution	C	Q	τ	τ'	t_w
1	5.738×10^5	4.861×10^{-2}	14.5	5.0	14.5
2	5.740×10^5	4.781×10^{-2}	14.5	4.5	14.5
3	5.743×10^5	4.570×10^{-2}	14.0	5.0	14.0
4	5.744×10^5	4.489×10^{-2}	14.0	4.5	14.0
5	5.750×10^5	4.416×10^{-2}	14.0	4.0	14.0
6	5.752×10^5	4.285×10^{-2}	13.5	5.0	13.5
7	5.753×10^5	4.206×10^{-2}	13.5	4.5	13.5
8	5.758×10^5	4.128×10^{-2}	13.5	4.0	13.5
9	5.775×10^5	3.850×10^{-2}	13.0	4.0	13.0
10	5.793×10^5	3.744×10^{-2}	12.5	5.0	12.5
…					
50	7.862×10^5	1.073×10^{-23}	5.5	4.0	9.5
51	7.864×10^5	1.063×10^{-2}	5.5	4.0	9.0
52	7.865×10^5	1.052×10^{-2}	5.5	4.0	8.5
53	7.867×10^5	1.041×10^{-2}	5.5	4.0	8.0
54	7.870×10^5	1.031×10^{-2}	5.5	4.0	7.5
55	7.873×10^5	1.022×10^{-2}	5.5	4.0	7.0
56	7.878×10^5	1.015×10^{-2}	5.5	4.0	6.5
57	7.884×10^5	1.009×10^{-3}	5.5	4.0	6.0
58	7.888×10^5	9.885×10^{-3}	5.5	4.0	5.5
59	8.300×10^5	9.869×10^{-3}	5.0	4.0	10.5
60	8.301×10^5	9.789×10^{-3}	5.0	4.0	10.0
…					
84	9.532×10^5	8.068×10^{-3}	4.0	4.0	9.0
85	9.534×10^5	8.001×10^{-3}	4.0	4.0	8.5
86	9.537×10^5	7.942×10^{-3}	4.0	4.0	8.0
87	9.541×10^5	7.938×10^{-3}	4.0	4.0	7.5
88	9.545×10^5	7.870×10^{-3}	4.0	4.0	7.0
89	9.549×10^5	7.803×10^{-3}	4.0	4.0	6.5
90	9.554×10^5	7.733×10^{-3}	4.0	4.0	6.0
91	9.562×10^5	7.721×10^{-3}	4.0	4.0	5.5
92	9.570×10^5	7.655×10^{-3}	4.0	4.0	5.0
93	9.581×10^5	7.599×10^{-3}	4.0	4.0	4.5
94	9.594×10^5	7.554×10^{-3}	4.0	4.0	4.0

the non-dominated solutions. It can be seen that each cluster is made up of solutions characterized by the same value of τ. Solutions 1 and 2, for which $\tau = 14.5$, constitute the first, two-elements cluster in the top-left part of Fig. 7. Then, solutions 3, 4 and 5 for which $\tau = 14$ constitute the second, three-elements cluster and so on. Intuitively, the structure of the Pareto-optimal frontier suggests that, as expected, the two objective functions are more sensitive to changes in the value of τ, i.e. the main inspection interval, rather than of τ' and t_{w}. In fact, a change in τ causes the solution to change cluster, whereas a change in τ' or t_{w} just tunes the objective functions' values in their neighbourhood. The inspection interval τ assumes its maximum value of 14.5 in correspondence of solution 1, the best solution from the economical perspective. Then, for increasing costs of maintenance (i.e. decreasing breakage probability), τ decreases until it reaches a value of 4 in correspondence with solution 94, which is characterized by the highest value of the cost and the lowest rail breakage probability. Note that in correspondence with this solution all the three control variables assume the value 4, which is the lower extreme in all the three corresponding search intervals. This is a consequence of the fact that the rail breakage probability increases monotonically with the three variables τ, τ' and t_{w} as it should be since the longer the two inspection times or the waiting time are, the more probable the rail breakage is. On the contrary, from the economical perspective, the results are conflicting: too short values of τ, τ' and t_{w} lead to too high expenditures related to too frequent inspections and maintenance operations; on the other hand, too large values of τ, τ' and t_{w} lead to too high expenditures due to unplanned maintenance and accident costs. Thus, the minimum value of the cost is reached in correspondence with solution 1, which suggests the values of 14.5, 5 and 14.5 for τ, τ' and t_{w}, respectively.

Solution 1, in the top-left region of the Pareto optimal frontier, represents the optimal inspection and maintenance strategy with respect to the cost but it is also the solution with the highest breakage probability among those in the Pareto optimal set. Then, moving along the Pareto optimal frontier of Fig. 7, one finds alternative solutions which, though being progressively more costly, are characterized by a lower breakage probability, i.e. by a higher level of safety. Thus, given the non-dominated solutions, characterized by different levels of cost and safety, the decision-makers can select, according to their respective attitude, a risk-informed preferred maintenance strategy. For example, the steep slope of the top-left region of the Pareto optimal frontier of Fig. 7, suggests that shifting from solution 1 and moving downwards along the frontier, a decision maker could select a maintenance strategy characterized by a considerably higher level of safety, though sacrificing little in the economical competitiveness. For example, solution 10 in Table 8 is characterized by a decrease in associated rail failure probability of 23% and an increase in costs of 1%.

We conclude that adopting a multiobjective approach provides more information to the decision maker, without the need of introducing any 'a priori' arbitrariness in the setting of the constraints which, instead, can be brought into play 'a posteriori', at the decision level. Indeed, although more informative, Pareto optimality does not solve the decision problem. The decision maker is actually provided with the whole spectrum of performances with respect to the objectives, but he or she must ultimately select the preferred one according to some subjective preference values.

Hence, the aim of the multi-objective GAs search procedure is to screen out those solutions, which are characterized by non-optimal performances with respect to all objectives. In the present application, out of 32768 potential solutions, the algorithm identifies 94 candidates (the Pareto-optimal set) for further analysis. From these, the closure of the problem can rely on techniques of decision analysis such as utility theory, multi-attribute value theory or fuzzy decision making, to name a few. Indeed, the application of such techniques is effective on a reduced set of potential alternatives (e.g. that of Fig. 7), which the decision maker can evaluate with respect to the decision weights, rules or monetary values assigned to the objective functions. This decision-making process falls beyond the scope of this paper. Yet, for the sake of completeness, we adopt, as an example, a popular criterion for choosing a 'single best compromise solution' is the so called 'min-max method' [10]. Let $S \equiv (f_1, f_2, \ldots, f_{N_{\mathrm{f}}})$ denote a generic point on the N_{f}-dimensional Pareto surface, and $f_i^{\min}$ $i = 1, 2, \ldots, N_{\mathrm{f}}$, the minimum value of the i-th objective function on such surface. For each point S we calculate the relative deviations $z_i = (f_i - f_i^{\min})/(f_i^{\max} - f_i^{\min})$, $i = 1, 2, \ldots, N_{\mathrm{f}}$ and take as a representative value $z_{\mathrm{S}} = \max_i(f_i)$. By definition, the best compromise solution is the point S^{o} on the Pareto surface corresponding to the minimum z_{S}.

In Fig. 7, the best compromise solution S^{o} for our case study is identified by the symbol 'o'. In order to aid the understanding of the rationale behind the min-max method, Fig. 8 reports the relative deviations z_i $i = 1, 2$

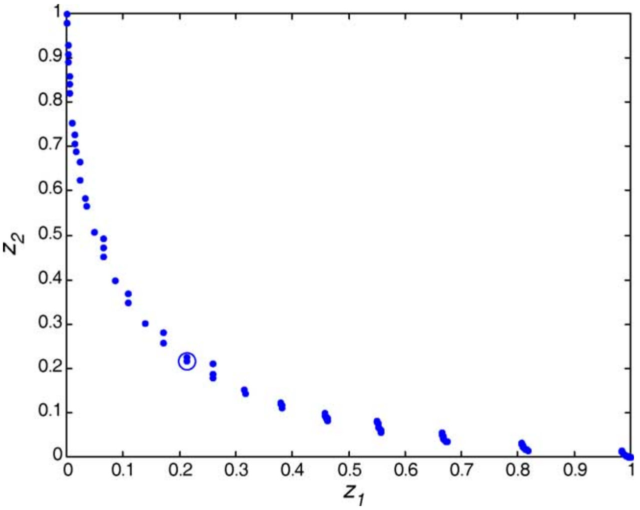


Fig. 8. Relative deviations of the Pareto solutions. The symbol 'o' indicates the best compromise solution according to the min-max method.

corresponding to each solution of the Pareto-optimal set of Fig. 7, with the best compromise solution S^o being again identified by 'o'. It can be seen that S^o is a best compromise in the sense that its 'goodness', measured in terms of the relative deviations z_i, $i=1,2$, is the most balanced between the two criteria of low cost, $f_1=C$, and breakage probability, $f_2=Q$: solutions located North-West of S^o have smaller relative deviation $z_1=(f_1-f_1^{\min})/(f_1^{\max}-f_1^{\min})$ and are, thus, unbalanced in favor of f_1; vice versa for solutions located South-East of S^o.

7. Conclusions

As a mean to re-launch competitiveness of the railway business over alternative forms of transportation, reliability-based and risk-informed approaches are nowadays flourishing within an effort of optimising railway infrastructures.

In this paper, we have developed a risk-informed methodology for the determination of an optimal inspection/maintenance strategy with the objective of reducing the operation and maintenance expenditures while still assuring high safety standards. Two sides of the problem have been considered in this work. On one side, a risk/cost model has been built accounting, as closely as possible, for the realistic issues of the rail failure process and for the actual inspection and maintenance procedures followed by the railway company. On the other side, a multiobjective optimisation viewpoint has been adopted addressing both the economical and safety-related aspects of railway operation. The multi-objective search has been performed by genetic algorithms, which have proven efficient in their task.

The multiobjective approach provides the decision makers with a spectrum of non-dominated solutions, characterized by different levels of cost and safety, within which a risk-informed preferred maintenance strategy should be selected, according to the decision makers respective attitude towards the two objectives. The multiobjective approach turns out to relieve the decision maker from the setting of any 'a priori' preference weight or constraint which, instead, can come into play 'a posteriori', at the decision level. Thus, if the decision maker were to change his or her constraints or preferences, these can be applied directly on the Pareto results of the multiobjective optimisation without having to run again the whole optimisation procedure. In alternative, as described in this paper, compromise solutions can be sought automatically, for example by using the popular 'min–max method'.

The model and approach proposed are intended to realistically inform decision-making on inspection and maintenance procedures in the railway industry. For this reason, the level of detail of the model has been refined to reproduce the actual procedures as closely as reasonable with respect to data availability. In order to estimate the model parameters, generic data was used from railway statistics, literature and expert judgement. Data collection studies are currently being undertaken by the Norwegian National Rail Administration in order to refine the assessment of cost figures and degradation parameters. Results from this study are expected to foster the applicability of the approach in practice.

References

[1] Carretero J, Pérez JM, García-Carballeira F, Calderón A, Fernández J, García JD, Lozano A, Cardona L, Cotaina N, Prete P, Applying RCM. in large scale systems: a case study with railway networks. Reliab Eng Syst Safety 2003;82:257–73.

[2] REMAIN consortium. Final consolidated progress report, European Union February; 1998

[3] Vatn J, Hokstad P, Bodsberg L. An overall model for maintenance optimization. Reliab Eng Syst Safety 1996;51:241–57.

[4] Svee H, Sæbø HJ, Vatn J. Estimating the potential benefit of introducing reliability centered maintenance on railway infrastructures. SINTEF technical report; 2000.

[5] Backer RD, Christer AH. Review of delay-time OR modelling of engineering aspects of maintanannce. Eur J Oper Res 1994;73:407–22.

[6] Zarembski AM, Palese JW. Risk based ultrasonic rail test scheduling: practical applications in Europe and North America. Sixth international conference on contact mechanics and wear of rail/wheel systems (CM2003) in Gothenburg, Sweden June 10–13; 2003, pp. 357–68.

[7] Vatn J, Svee H. A risk based approach to determine ultrasonic inspection frequencies in railway applications. JBV Technical report; 2000.

[8] Podofillini L, Zio E, Vatn J. Modelling the degrading failure of a rail section under periodic inspection. PSAM 7/ESREL, Berlin, Germany, June 14–18 2004;2570–5.

[9] Bukowski JV. Modelling and analyzing the effects of periodic inspection on the performance of safety critical systems. IEEE Trans Reliab 2001;50(3):321–9.

[10] Belegundu AD, Chandrupatla TR. Optimization concepts and applications in engineering, Prentice Hall Editions; chapter 11, pp. 373–81

[11] Holland JH. Adaptation in natural and artificial system. Ann Arbor, MI: University of Michigan Press; 1975.

[12] Goldberg DE. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley; 1989.

[13] Beasly D, Bull D, Martin R. An overview on genetic algorithms: Part 1 fundamentals. Univ Comput 1993;15(2):58–69.

[14] Beasly D, Bull D, Martin R. An overview on genetic algorithms: Part 2 research topics. Univ Comput 1993;15(4):170–81.

[15] Joyce PA, Withers TA, Hickling PJ. Application of Genetic Algorithms to Optimum Offshore Plant Design. Proceedings of ESREL 98, Trondheim (Norway), June 16–19 1998 pp. 665–71.

[16] Painton L, Campbell J. Genetic algorithms in optimization of system reliability. IEEE Trans Reliab 1995;44(2):172–8.

[17] Coit DW, Smith AE. Reliability optimization of series-parallel systems using a genetic algorithm. IEEE Trans Reliab 1996;45(2):254–60.

[18] Martorell S, Carlos S, Sanchez A, Serradell V. Constrained optimization of test intervals using a steady-state genetic algorithm. Reliab Engng Sys Safety 2000;67:215–32.

[19] Marseguerra M, Zio E. Optimizing maintenance and repair policies via a combination of genetic algorithms and Monte Carlo simulation. Reliab Eng Syst Safety 2000;68:69–83.

[20] Sawaragi Y, Nakayama H, Tanino T. Theory of multiobjective optimization. Orlando, FL: Academic Press; 1985.

[21] Zitzler E, Thiele L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. IEEE Trans Evol Comput 1999;3(4):257–71.

[22] Zitzler E. Evolutionary algorithms for multiobjective optimization: methods and applications. Swiss Federal Institute of Technology (ETH) Zurich. TIK-Schriftenreihe Nr. 30, Diss ETH No. 13398. Germany: Shaker Verlag; 1999. ISBN 3-8265-6831-1.

[23] Parks GT. Multiobjective pressurized water reactor reload core design using genetic algorithm search. Nucl Sci Eng 1997;124:178–87.

[24] Toshinsky VG, Sekimoto H, Toshinsky GI. A method to improve multiobjective genetic algorithm optimization of a self-fuel-providing LMFBR by niche induction among nondominated solutions. Ann Nucl Energy 2000;27:397–410.

[25] Giuggioli Busacca P, Marseguerra M, Zio E. Multiobjective optimization by genetic algorithm: application to safety systems. Reliab Eng Syst Safety 2001;72:59–74.

[26] Papazoglou IA. Semi-Markovian reliability models for systems with testable components and general test/outage times. Reliab Eng Syst Safety 2000;68:121–33.